Device Drivers, Features, and Commands on Red Hat Enterprise Linux 7.5
Device Drivers, Features, and Commands on Red Hat Enterprise Linux 7.5
# Contents

Summary of changes ........................................ vii
About this publication ..................................... ix

**Part 1. General concepts** .................................. 1

Chapter 1. How devices are accessed by Linux. ................. 3
Chapter 2. Devices in sysfs .................................. 7
Chapter 3. Kernel and module parameters ...................... 23

**Part 2. Booting and shutdown** .............................. 31

Chapter 4. Console device drivers ........................... 33
Chapter 5. Initial program loader for IBM Z - zipl. ........... 55
Chapter 6. Booting Linux .................................... 81
Chapter 7. Suspending and resuming Linux .................... 103
Chapter 8. Shutdown actions ................................ 109
Chapter 9. The diag288 watchdog device driver. ............... 115

**Part 3. Storage** ........................................... 119

Chapter 10. DASD device driver .............................. 121
Chapter 11. SCSI-over-Fibre Channel device driver ............ 159
Chapter 12. Storage-class memory device driver .............. 201
Chapter 13. Channel-attached tape device driver .............. 205
Chapter 14. XPRAM device driver ............................ 215

**Part 4. Networking** ....................................... 219

Chapter 15. qeth device driver for OSA-Express (QDIO) and HiperSockets .... 223
Chapter 16. OSA-Express SNMP subagent support ............... 289
Chapter 17. LAN channel station device driver ................. 299
Chapter 18. CTCM device driver .............................. 305

© Copyright IBM Corp. 2000, 2018
Chapter 19. AF_IUCV address family support ........................................... 319
Chapter 20. RDMA over Converged Ethernet ........................................... 323

Part 5. System resources ................................................................. 325
Chapter 21. Managing CPUs ............................................................. 327
Chapter 22. Managing hotplug memory ............................................... 333
Chapter 23. Large page support ......................................................... 339
Chapter 24. S/390 hypervisor file system ............................................ 343
Chapter 25. ETR- and STP-based clock synchronization ........................... 349
Chapter 26. Identifying the IBM Z hardware ......................................... 353
Chapter 27. HMC media device driver ................................................ 355
Chapter 28. Data compression with GenWQE and zEDC Express ............... 359

Part 6. z/VM virtual server integration .............................................. 369
Chapter 29. z/VM concepts .............................................................. 371
Chapter 30. Writing kernel APPLDATA records ..................................... 375
Chapter 31. Writing z/VM monitor records .......................................... 381
Chapter 32. Reading z/VM monitor records ......................................... 385
Chapter 33. z/VM recording device driver ............................................ 391
Chapter 34. z/VM unit record device driver ......................................... 399
Chapter 35. z/VM DCSS device driver ................................................. 401
Chapter 36. z/VM CP interface device driver ....................................... 413
Chapter 37. z/VM special messages uevent support ................................. 415
Chapter 38. Cooperative memory management .................................... 421

Part 7. Security ............................................................................ 423
Chapter 39. Generic cryptographic device driver .................................... 425
Chapter 40. Pseudorandom number generator device driver ..................... 441
Chapter 41. True random-number generator device driver ....................... 445
Chapter 42. Protected key device driver ............................................. 447
<table>
<thead>
<tr>
<th>Chapter 43. Hardware-accelerated in-kernel cryptography</th>
<th>449</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 44. Instruction execution protection</td>
<td>453</td>
</tr>
<tr>
<td>Part 8. Performance measurement using hardware facilities.</td>
<td>455</td>
</tr>
<tr>
<td>Chapter 45. Channel measurement facility</td>
<td>457</td>
</tr>
<tr>
<td>Chapter 46. OProfile hardware sampling support</td>
<td>461</td>
</tr>
<tr>
<td>Chapter 47. Using the CPU-measurement facilities</td>
<td>465</td>
</tr>
<tr>
<td>Part 9. Diagnostics and troubleshooting</td>
<td>471</td>
</tr>
<tr>
<td>Chapter 48. Logging I/O subchannel status information</td>
<td>473</td>
</tr>
<tr>
<td>Chapter 49. Control program identification</td>
<td>475</td>
</tr>
<tr>
<td>Chapter 50. Activating automatic problem reporting</td>
<td>479</td>
</tr>
<tr>
<td>Chapter 51. Check whether the Linux instance can be a hypervisor</td>
<td>481</td>
</tr>
<tr>
<td>Chapter 52. Avoiding common pitfalls</td>
<td>483</td>
</tr>
<tr>
<td>Part 10. Reference</td>
<td>487</td>
</tr>
<tr>
<td>Chapter 53. Commands for Linux on Z</td>
<td>489</td>
</tr>
<tr>
<td>Chapter 54. Selected kernel parameters</td>
<td>651</td>
</tr>
<tr>
<td>Chapter 55. Linux diagnose code use</td>
<td>671</td>
</tr>
<tr>
<td>Part 11. Appendixes</td>
<td>673</td>
</tr>
<tr>
<td>Appendix A. Accessibility</td>
<td>675</td>
</tr>
<tr>
<td>Appendix B. Understanding syntax diagrams.</td>
<td>677</td>
</tr>
<tr>
<td>Notices</td>
<td>679</td>
</tr>
<tr>
<td>Glossary</td>
<td>681</td>
</tr>
<tr>
<td>Bibliography</td>
<td>687</td>
</tr>
<tr>
<td>Index</td>
<td>691</td>
</tr>
</tbody>
</table>
Summary of changes

Updates for Red Hat Enterprise Linux 7.5

This update contains the following changes.

New information

- The cryptographic device driver now supports multiple domains, see “Cryptographic domains” on page 428 and “Displaying information about cryptographic devices” on page 432.
- A new device driver provides true random numbers to applications in user space. See Chapter 41, “True random-number generator device driver,” on page 445.
- With a new device driver, cryptographic applications can generate protected keys from secure keys or from clear keys. Protected keys can be used by CPACF for accelerated encryption and decryption. See Chapter 42, “Protected key device driver,” on page 447.
- You can now use the z/Architecture® instruction execution protection feature, see Chapter 44, “Instruction execution protection,” on page 453.
- There is a new command to generate, validate, and re-encipher secure AES keys. See “zkey - Generate, validate, and re-encipher secure AES keys” on page 645.

Changed Information

- None.

This revision also includes maintenance and editorial changes. Technical changes or additions to the text and illustrations are indicated by a vertical line to the left of the change.

Deleted Information

- None.

Updates for Red Hat Enterprise Linux 7.3 and 7.4

This edition contains changes related to release 7.4 of Red Hat Enterprise Linux.

New information

- Linux now supports UIDs as persistent identifiers for PCI functions, see “PCI Express support” on page 19.
- You can now read measurement data for PCIe devices, see “Reading statistics for a PCIe device” on page 22.
- Channel paths that are subject to frequent IFCC or CCC errors can now be taken offline automatically, see “Setting defective channel paths offline automatically” on page 153.
- You can now use 2 GB large pages when Linux is running on an LPAR, see Chapter 23, “Large page support,” on page 339.
- The GenWQE device driver facilitates hardware-accelerated data compression and decompression through a PCIe-attached Field Programmable Gate Array (FPGA) acceleration adapter, see Chapter 28, “Data compression with GenWQE and zEDC Express,” on page 359.
• Information has been added about using hardware-acceleration for in-kernel cryptographic operations, see Chapter 43, “Hardware-accelerated in-kernel cryptography,” on page 449.

• You can now check whether a Linux instance can be a hypervisor, see Chapter 51, “Check whether the Linux instance can be a hypervisor,” on page 481.

• The `dasdfmt` command now offers a quick format mode for DASD that have previously been formatted with the cd1 or ld1 disk layout, see “dasdfmt - Format a DASD” on page 520.

• With new parameters for the `vmur` command, you can control the CLASS, DEST, FORM, and DIST spooling options for virtual unit record devices. See “vmur - Work with z/VM spool file queues” on page 631.

• The `dasdfmt` command and the `zdsfs` commands now check whether a DASD volume is online to another operating system instance, see “dasdfmt - Format a DASD” on page 520 and “zdsfs - Mount a z/OS DASD” on page 640.

• A new section describes the `fips` kernel parameter, which enables the FIPS mode of operation, “fips - Run Linux in FIPS mode” on page 657.

**Changed Information**

• You can now IPL from subchannel sets other than 0, see “Attributes for ccw” on page 100.

• The CPU topology information now includes drawers, see “Examining the CPU topology” on page 330.

This revision also includes maintenance and editorial changes. Technical changes or additions to the text and illustrations are indicated by a vertical line to the left of the change.

**Deleted Information**

• None.
About this publication

This publication describes the device drivers, features, and commands available to Red Hat Enterprise Linux 7.5 for the control of IBM® Z devices and attachments. Unless stated otherwise, in this book the terms device drivers and features are understood to refer to device drivers and features for Red Hat Enterprise Linux 7.5 for z Systems.

Unless stated otherwise, all z/VM® related information in this document assumes a current z/VM version, see [www.vm.ibm.com/techinfo/lpmigr/vmleos.html](http://www.vm.ibm.com/techinfo/lpmigr/vmleos.html).

As of July 2017, IBM z Systems® is re-branded to IBM Z. In this document, Linux on z Systems refers to Linux running on LinuxONE or an IBM mainframe, including all IBM mainframe systems supported by Red Hat Enterprise Linux. In particular, this includes IBM z14 (z14), IBM z13® (z13), IBM zEnterprise® BC12 (zBC12), IBM zEnterprise EC12 (zEC12), IBM zEnterprise 196 (z196), IBM zEnterprise 114 (z114).

For a support matrix, see the Capabilities and Limits section in the Red Hat Enterprise Linux 7.5 release notes at [https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux](https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux). The same Web page contains Technical Notes with details on various features and known issues.

You can find the latest version of this publication on the developerWorks® website at [www.ibm.com/developerworks/linux/linux390/documentation_red_hat.html](http://www.ibm.com/developerworks/linux/linux390/documentation_red_hat.html).

How this publication is organized

The first part of this publication contains general and overview information for the z/Architecture specific device drivers for Red Hat Enterprise Linux 7.5 for z Systems.

Part two contains chapters about device drivers and features that are used in the context of booting and shutting down Linux.

Part three contains chapters specific to individual storage device drivers.

Part four contains chapters specific to individual network device drivers.

Part five contains chapters about device drivers and features that help to manage the resources of the real or virtual hardware.

Part six contains chapters that describe device drivers and features in support of z/VM virtual server integration.

Part seven contains chapters about device drivers and features that support security aspects of Red Hat Enterprise Linux 7.5 for z Systems.

Part eight contains chapters about assessing the performance of Red Hat Enterprise Linux 7.5 for z Systems.
Part nine contains chapters about device drivers and features that are used in the context of diagnostics and problem solving.

Part ten contains chapters with reference information about commands, kernel parameters, and Linux use of z/VM DIAG calls.

Who should read this publication

Most of the information in this publication is intended for system administrators who want to configure Red Hat Enterprise Linux 7.5 for z Systems.

The following general assumptions are made about your background knowledge:

- You have an understanding of basic computer architecture, operating systems, and programs.
- You have an understanding of Linux and IBM Z terminology.
- You are familiar with Linux device driver software.
- You are familiar with the Z devices attached to your system.

Programmers: Some sections are of interest primarily to specialists who want to program extensions to the device drivers and features.

Conventions and assumptions used in this publication

This section summarizes the styles, highlighting, and assumptions used throughout the publication.

Authority

Most of the tasks described in this document require a user with root authority. In particular, writing to procfs, and writing to most of the described sysfs attributes requires root authority.

Throughout this document, it is assumed that you have root authority.

Making changes persistent

This document describes how to change settings and options for mainframe computers in sysfs. In most cases, changes in sysfs are not persistent. If you need to make your changes persistent, see Red Hat Enterprise Linux 7.5 System Administrator’s Guide for details about the configuration files to use.

This document describes how to load modules with modprobe. Loading a module this way is not persistent across reboots. If you want to load your kernel modules automatically at boot time, see the section on persistent module loading in Red Hat Enterprise Linux 7.5 System Administrator’s Guide.

Terminology

In this document, the term booting is used for running boot loader code that loads the Linux operating system. IPL is used for issuing an IPL command to load boot loader code or a stand-alone dump utility. See also “IPL and booting” on page 81.
**sysfs and procfs**

In this document, the mount point for the virtual Linux file system sysfs is assumed to be `/sys`. Correspondingly, the mount point for procfs is assumed to be `/proc`.

**debugfs**

This document assumes that debugfs has been mounted at `/sys/kernel/debug`.

To mount debugfs, you can use this command:

```
# mount -t debugfs none /sys/kernel/debug
```

**Documentation directory**

This document sometimes refers to files in the Documentation directory in the Linux source tree.

On Red Hat Enterprise Linux 7.5 the full path to this directory is:

```
/usr/share/doc/kernel-doc-<version>/Documentation
```

If this directory is not present, install the `kernel-doc-<version>.el7`.noarch RPM.

**Number prefixes**

In this publication, KB means 1024 bytes, MB means 1,048,576 bytes, and GB means 1,073,741,824 bytes.

**Hexadecimal numbers**

Mainframe documents and Linux documents tend to use different styles for writing hexadecimal numbers. Thirty-one, for example, would typically read X'1F' in a mainframe book and 0x1f in a Linux book.

Because the Linux style is required in many commands and is also used in some code samples, the Linux style is used throughout this publication.

**Highlighting**

This document uses the following highlighting styles:

- Paths and URLs are highlighted in **monospace**.
- Variables are highlighted in `<italics within angled brackets>`.
- Commands in text are highlighted in **monospace bold**.
- Input and output as normally seen on a computer screen is shown within a **screen frame**.
- Prompts are shown as **hash signs**: `#`
Part 1. General concepts

Chapter 1. How devices are accessed by Linux

Device names, device nodes, and major/minor numbers
Network interfaces

3

Chapter 2. Devices in sysfs

Device categories
Device directories
Device views in sysfs

7

Chapter 3. Kernel and module parameters

Specifying kernel parameters
Specifying parameters for modules

23

11

This information at an overview level describes concepts that apply across different device drivers and kernel features.

Newest version

You can find the newest version of this book at

www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_distlibs.html

© Copyright IBM Corp. 2000, 2018
Chapter 1. How devices are accessed by Linux

Applications on Linux access character and block devices through device nodes, and network devices through network interfaces.

Device names, device nodes, and major/minor numbers

The Linux kernel represents character and block devices as pairs of numbers \(<major>:<minor>\).

Some major numbers are reserved for particular device drivers. Other major numbers are dynamically assigned to a device driver when Linux boots. For example, major number 94 is always the major number for DASD devices while the device driver for channel-attached tape devices has no fixed major number. A major number can also be shared by multiple device drivers. See /proc/devices to find out how major numbers are assigned on a running Linux instance.

The device driver uses the minor number \(<minor>\) to distinguish individual physical or logical devices. For example, the DASD device driver assigns four minor numbers to each DASD: one to the DASD as a whole and the other three for up to three partitions.

Device drivers assign device names to their devices, according to a device driver-specific naming scheme (see, for example, “DASD naming scheme” on page 127). Each device name is associated with a minor number (see Figure 1).

![Figure 1. Minor numbers and device names](image)

User space programs access character and block devices through device nodes also referred to as device special files. When a device node is created, it is associated with a major and minor number (see Figure 2).

![Figure 2. Device nodes](image)
Red Hat Enterprise Linux 7.5 uses udev to create device nodes for you. There is always a device node that matches the device name that is used by the kernel, and additional nodes might be created by special udev rules. See the udev man page for more details.

**Network interfaces**

The Linux kernel representation of a network device is an interface.

![Diagram](image)

*Figure 3. Interfaces*

When a network device is defined, it is associated with a real or virtual network adapter (see Figure 3). You can configure the adapter properties for a particular network device through the device representation in sysfs (see “Device directories” on page 8).

You activate or deactivate a connection by addressing the interface with `ip` or an equivalent command. All interfaces that are provided by the IBM Z specific network device drivers are interfaces for the Internet Protocol (IP).

On Red Hat Enterprise Linux, you configure network interfaces through the NetworkManager. Use the `nmcli` command line tool to interact with the NetworkManager.

**Predictable network device names**

Red Hat Enterprise Linux 7 uses predictable names for network devices. These names are stable across reboots and network adapter replacements.

Predictable naming is enabled by default. In this naming scheme, a mainframe network device has a name of the following form:

```plaintext
<pf><type><bus_id>
```

For example:

```
enccw0.0.f5f0
```

Where:

- **<pf>** A two-character prefix for the network type. The type can be one of the following:
  - `en` - Ethernet
  - `ww` - WAN
  - `sl` - serial line, such as CTC

- **<type>** The device type. The device type of CCW devices is `ccw`.

- **<bus_id>** The bus ID identifies the device within the scope of a Linux instance. For a CCW device, the bus ID is the device number of the device with a leading “0.n.”, where n is the subchannel set ID, for example, 0.1.0ab1.
You can use `znetconf -c` or `lsccs -a` to obtain a list of the devices in your system.

For more information about the predictable naming scheme, see the Red Hat Enterprise Linux 7.5 Networking Guide available at

https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux

**Matching devices with the corresponding interfaces**

If you define multiple interfaces on a Linux instance, you must keep track of the interface names assigned to your network devices.

Red Hat Enterprise Linux 7.5 uses predictable network names, which make it easy to match network devices with their interface names. The interface names are preserved across reboots.

How you can list interface names and devices depends on the network device driver. For qeth, you can use the `lsqeth` command (see “lsqeth - List qeth-based network devices” on page 581) to obtain a mapping. To see the mapping for CTC devices also, use `znetconf` (see “znetconf - List and configure network devices” on page 648).

After setting a device online, read `/var/log/messages` or issue `journalctl` to find the associated interface name in the messages that are issued in response to the device being set online.

For each network device that is online, there is a symbolic link of the form `/sys/class/net/<interface>/device` where `<interface>` is the interface name. This link points to a sysfs directory that represents the corresponding network device. You can read this symbolic link with `readlink` to confirm that an interface name corresponds to a particular network device.

“Device views in sysfs” on page 11 tells you where you can find the device directories with their attributes in sysfs.

**Main steps for setting up a network interface**

The main steps apply to all network devices drivers that are based on ccwgroup devices (for example, qeth and lcs devices). How to perform a particular step can be different for the different device drivers.

The main steps are:

1. Create a network device by combining suitable subchannels into a group device. The device driver then creates directories that represent the device in sysfs.

2. Configure the device through its attributes in sysfs. See “Device views in sysfs” on page 11. Some devices have attributes that can or must be set later when the device is online or when the connection is active.

3. Set the device online. This step associates the device with an interface name and thus makes the device known to the Linux network stack. For devices that are associated with a physical network adapter it also initializes the adapter for the network interface.

4. Configure and activate the interface. This step adds interface properties like IP addresses, netmasks, and MTU to the network interface and moves the network interface into state "up". The interface is then ready for user space (socket) programs to run connections and transfer data across it.
To configure a network device, use tools provided with Red Hat Enterprise Linux. See Red Hat Enterprise Linux 7.5 Networking Guide
Chapter 2. Devices in sysfs

Most Linux device drivers create structures in sysfs. These structures hold information about individual devices and are also used to configure and control the devices.

Device categories

The /sys/devices directory includes several device categories that are specific to z/Architecture.

Figure 4 illustrates a part of sysfs.

Figure 4. sysfs

/sys/bus and /sys/devices are common Linux directories. The directories following /sys/bus sort the device drivers according to the categories of devices they control. The sysfs branch for a particular category might be missing if there is no device for that category.

AP devices are adjunct processors used for cryptographic operations.

CCW devices are devices that can be addressed with channel-command words (CCWs). These devices use a single subchannel on the mainframe’s channel subsystem.

CCW group devices are devices that use multiple subchannels on the mainframe’s channel subsystem.
IUCV devices
are devices for virtual connections between z/VM guest virtual machines
within an IBM mainframe. IUCV devices do not use the channel subsystem.

Table 1 lists the z/Architecture specific device drivers that have representation in sysfs:

<table>
<thead>
<tr>
<th>Device driver</th>
<th>Category</th>
<th>sysfs directories</th>
</tr>
</thead>
<tbody>
<tr>
<td>3215 console</td>
<td>CCW</td>
<td>/sys/bus/ccw/drivers/3215</td>
</tr>
<tr>
<td>3270 console</td>
<td>CCW</td>
<td>/sys/bus/ccw/drivers/3270</td>
</tr>
<tr>
<td>DASD</td>
<td>CCW</td>
<td>/sys/bus/ccw/drivers/dasd-eckd</td>
</tr>
<tr>
<td></td>
<td></td>
<td>/sys/bus/ccw/drivers/dasd-fba</td>
</tr>
<tr>
<td>SCSI-over-Fibre Channel</td>
<td>CCW</td>
<td>/sys/bus/ccw/drivers/zfcp</td>
</tr>
<tr>
<td>Channel-attached tape</td>
<td>CCW</td>
<td>/sys/bus/ccw/drivers/tape_34xx</td>
</tr>
<tr>
<td></td>
<td></td>
<td>/sys/bus/ccw/drivers/tape_3590</td>
</tr>
<tr>
<td>Cryptographic</td>
<td>AP</td>
<td>/sys/bus/ap/drivers/cex3a</td>
</tr>
<tr>
<td>DCSS</td>
<td>n/a</td>
<td>/sys/devices/dcssblk</td>
</tr>
<tr>
<td>XPRAM</td>
<td>n/a</td>
<td>/sys/devices/system/xpram</td>
</tr>
<tr>
<td>z/VM recording</td>
<td>IUCV</td>
<td>/sys/bus/iucv/drivers/vmlogrdr</td>
</tr>
<tr>
<td>qeth (OSA-Express features and HiperSockets™)</td>
<td>CCW group</td>
<td>/sys/bus/ccwgroup/drivers/qeth</td>
</tr>
<tr>
<td>LCS</td>
<td>CCW group</td>
<td>/sys/bus/ccwgroup/drivers/lcs</td>
</tr>
<tr>
<td>CTCM</td>
<td>CCW group</td>
<td>/sys/bus/ccwgroup/drivers/ctcm</td>
</tr>
</tbody>
</table>

Some device drivers do not relate to physical devices that are connected through the channel subsystem. Their representation in sysfs differs from the CCW and CCW group devices, for example, the Cryptographic device drivers have their own category, AP.

The following sections provide more details about devices and their representation in sysfs.

Device directories

Each device that is known to Linux is represented by a directory in sysfs.

For CCW and CCW group devices the name of the directory is a bus ID that identifies the device within the scope of a Linux instance. For a CCW device, the bus ID is the device's device number with a leading “0.<n>.”, where <n> is the subchannel set ID. For example, 0.1.0ab1.

CCW group devices are associated with multiple device numbers. For CCW group devices, the bus ID is the primary device number with a leading “0.<n>.”, where <n> is the subchannel set ID.

“Device views in sysfs” on page 11 tells you where you can find the device directories with their attributes in sysfs. Red Hat Enterprise Linux 7.5 uses configuration files to control devices. For example, network devices have interface
scripts called /etc/sysconfig/network-scripts/ifcfg-<interface-name>. See the Red Hat Enterprise Linux 7.5 Networking Guide for details about configuration files.

Device attributes

The device directories contain attributes. You control a device by writing values to its attributes.

Some attributes are common to all devices in a device category, other attributes are specific to a particular device driver. The following attributes are common to all CCW devices:

**online**
You use this attribute to set the device online or offline. To set a device online, write the value 1 to its online attribute. To set a device offline, write the value 0 to its online attribute.

**cutype**
specifies the control unit type and model, if applicable. This attribute is read-only.

**cmb_enable**
Enables I/O data collection for the device. See "Enabling, resetting, and switching off data collection” on page 458 for details.

**devtype**
specifies the device type and model, if applicable. This attribute is read-only.

**availability**
indicates whether the device can be used. The following values are possible:

**good**
This is the normal state. The device can be used.

**boxed**
The device is locked by another operating system instance and cannot be used until the lock is surrendered or the DASD is accessed by force (see "Accessing DASD by force” on page 137).

**no device**
Applies to disconnected devices only. The device disappears after a machine check and the device driver requests to keep the device online anyway. Changes back to “good” when the device returns after another machine check and the device driver accepts the device back.

**no path**
Applies to disconnected devices only. After a machine check or a logical vary off, no path remains to the device. However, the device driver keeps the device online. Changes back to “good” when the path returns after another machine check or logical vary on and the device driver accepts the device back.

**modalias**
contains the module alias for the device. It is of the format:

ccw:t<cu_type>m<cu_model>

or

ccw:t<cu_type>m<cu_model>d<dev_type>m<dev_model>
Setting attributes

Directly write to attributes or, for CCW devices, use the `chccwdev` command to set attribute values.

Procedure

- You can set a writable attribute by writing the designated value to the corresponding attribute file.
- For CCW devices, you can also use the `chccwdev` command (see "chccwdev - Set CCW device attributes" on page 490) to set attributes.

With a single `chccwdev` command you can:
- Set an attribute for multiple devices
- Set multiple attributes for a device, including setting the device online
- Set multiple attributes for multiple devices

Working with newly available devices

Errors can occur if you try to work with a device before its sysfs representation is completely initialized.

About this task

When new devices become available to a running Linux instance, some time elapses until the corresponding device directories and their attributes are created in sysfs. Errors can occur if you attempt to work with a device for which the sysfs structures are not present or are not complete. These errors are most likely to occur and most difficult to handle when you are configuring devices with scripts.

Procedure

Use the following steps before you work with a newly available device to avoid such errors:

1. Attach the device, for example, with a z/VM CP ATTACH command.
2. Assure that the sysfs structures for the new device are complete.
   
   ```
   # echo 1 > /proc/cio_settle
   ```

   This command returns control after all pending updates to sysfs are complete.

   **Tip:** For CCW devices, you can omit this step if you then use `chccwdev` (see "chccwdev - Set CCW device attributes" on page 490) to work with the devices.

   `chccwdev` triggers cio_settle for you and waits for cio_settle to complete.

3. Assure that udev actions for the new device are complete.
   
   ```
   # udevadm settle
   ```

   The `settle` command returns control after all pending events are complete.

Results

You can now work with the new device. For example, you can set the device online or set attributes for the device.
Device views in sysfs

sysfs provides multiple views of device specific data.

The most important views are:

- "Device driver view"
- "Device category view"
- "Device view" on page 12
- "Channel subsystem view" on page 12

Many paths in sysfs contain device bus-IDs to identify devices. Device bus-IDs of subchannel-attached devices are of the form:

0.<n>.<devno>

where <n> is the subchannel set-ID and <devno> is the device number.

Device driver view

This view groups devices by the device drivers that control them.

The device driver view is of the form:

/sys/bus/<bus>/drivers/<driver>/<device_bus_id>

where:

<bus> is the device category, for example, ccw or ccwgroup.
<driver> is a name that specifies an individual device driver or the device driver component that controls the device (see Table 1 on page 8).
<device_bus_id> identifies an individual device (see "Device directories" on page 8).

Note: DCSSs and XPRAM are not represented in this view.

Examples

- This example shows the path for an ECKD™ type DASD device:
  /sys/bus/ccw/drivers/dasd-eckd/0.0.b100
- This example shows the path for a qeth device:
  /sys/bus/ccwgroup/drivers/qeth/0.0.a100
- This example shows the path for a cryptographic device (a CEX3A card):
  /sys/bus/ap/drivers/cex3a/card3b

Device category view

This view groups devices by major categories that can span multiple device drivers.

The device category view does not sort the devices according to their device drivers. All devices of the same category are contained in a single directory. The device category view is of the form:

/sys/bus/<bus>/devices/<device_bus_id>

where:
<bus> is the device category, for example, ccw or ccwgroup.

<device_bus_id> identifies an individual device (see “Device directories” on page 8).

Note: DCSSs and XPRAM are not represented in this view.

Examples

- This example shows the path for a CCW device.
  /sys/bus/ccw/devices/0.0.b100

- This example shows the path for a CCW group device.
  /sys/bus/ccwgroup/devices/0.0.a100

- This example shows the path for a cryptographic device:
  /sys/bus/ap/devices/card3b

Device view

This view sorts devices according to their device drivers, but independent from the device category. It also includes logical devices that are not categorized.

The device view is of the form:
/sys/devices/<driver>/<device>

where:

<driver>
  is a name that specifies an individual device driver or the device driver component that controls the device.

<device>
  identifies an individual device. The name of this directory can be a device bus-ID or the name of a DCSS or IUCV device.

Examples

- This example shows the path for a qeth device.
  /sys/devices/qeth/0.0.a100

- This example shows the path for a DCSS block device.
  /sys/devices/dcssblk/mydcss

Channel subsystem view

The channel subsystem view shows the relationship between subchannels and devices.

The channel subsystem view is of the form:
/sys/devices/css0/<subchannel>

where:

<subchannel>
  is a subchannel number with a leading “0.<n>.”, where <n> is the subchannel set ID.

I/O subchannels show the devices in relation to their respective subchannel sets and subchannels. An I/O subchannel is of the form:
/sys/devices/css0/<subchannel>/<device_bus_id>
where:

<subchannel>

is a subchannel number with a leading “0.<n>.”, where <n> is the subchannel set ID.

<device_bus_id>

is a device number with a leading “0.<n>.”, where <n> is the subchannel set ID (see “Device directories” on page 8).

Examples

- This example shows a CCW device with device number 0xb100 that is associated with a subchannel 0x0001.
  
  /sys/devices/css0/0.0.0001/0.0.b100

- This example shows a CCW device with device number 0xb200 that is associated with a subchannel 0x0001 in subchannel set 1.
  
  /sys/devices/css0/0.1.0001/0.1.b200

- The entries for a group device show as separate subchannels. If a CCW group device uses three subchannels 0x0002, 0x0003, and 0x0004 the subchannel information could be:
  
  /sys/devices/css0/0.0.0002/0.0.a100
  /sys/devices/css0/0.0.0003/0.0.a101
  /sys/devices/css0/0.0.0004/0.0.a102

Each subchannel is associated with a device number. Only the primary device number is used for the bus ID of the device in the device driver view and the device view.

- This example lists the information available for a non-I/O subchannel with which no device is associated:
  
  ls /sys/devices/css0/0.0.ff00/
  bus driver modalias subsystem type uevent

Subchannel attributes

There are sysfs attributes that represent subchannel properties, including common attributes and information specific to the subchannel type.

Subchannels have two common attributes:

**type**

The subchannel type, which is a numerical value, for example:

- 0 for an I/O subchannel
- 1 for a CHSC subchannel
- 3 for an EADM subchannel

**modalias**

The module alias for the device of the form css:t<n>, where <n> is the subchannel type (for example, 0 or 1).

These two attributes are the only ones that are always present. Some subchannels, like I/O subchannels, might contain devices and further attributes.

Apart from the bus ID of the attached device, I/O subchannel directories typically contain these attributes:
chpids
is a list of the channel-path identifiers (CHPID) through with the device is connected. See also “Channel path ID information” on page 15.

pimpampom
provides the path installed, path available, and path operational masks. See z/Architecture Principles of Operation, SA22-7832 for details about the masks.

Channel path measurement

A sysfs attribute controls the channel path measurement facility of the channel subsystem.

/sys/devices/css0/cm_enable

With the cm_enable attribute you can enable and disable the extended channel-path measurement facility. It can take the following values:

0  Deactivates the measurement facility and remove the measurement-related attributes for the channel paths. No action if measurements are not active.

1  Attempts to activate the measurement facility and create the measurement-related attributes for the channel paths. No action if measurements are already active.

If a machine does not support extended channel-path measurements the cm_enable attribute is not created.

Two sysfs attributes are added for each channel path object:

cmg  Specifies the channel measurement group or unknown if no characteristics are available.

shared  Specifies whether the channel path is shared between LPARs or unknown if no characteristics are available.

If measurements are active, two more sysfs attributes are created for each channel path object:

measurement  A binary sysfs attribute that contains the extended channel-path measurement data for the channel path. It consists of eight 32-bit values and must always be read in its entirety, or 0 will be returned.

measurement_chars  A binary sysfs attribute that is either empty, or contains the channel measurement group dependent characteristics for the channel path, if the channel measurement group is 2 or 3. If not empty, it consists of five 32-bit values.

Examples

• To turn measurements on issue:
  
  # echo 1 > /sys/devices/css0/cm_enable

• To turn measurements off issue:
  
  # echo 0 > /sys/devices/css0/cm_enable
Channel path ID information

All CHPIDs that are known to Linux are shown alongside the subchannels in the /sys/devices/css0 directory.

The directories that represent the CHPIDs have the form:
/sys/devices/css0/chp0.<chpid>

where <chpid> is a two digit hexadecimal CHPID.

Example: /sys/devices/css0/chp0.4a

Setting a CHPID logically online or offline

Directories that represent CHPIDs contain a status attribute that you can use to set the CHPID logically online or offline.

About this task

When a CHPID has been set logically offline from a particular Linux instance, the CHPID is, in effect, offline for this Linux instance. A CHPID that is shared by multiple operating system instances can be logically online to some instances and offline to others. A CHPID can also be logically online to Linux while it has been varied off at the SE.

Procedure

To set a CHPID logically online, set its status attribute to online by writing the value on to it. To set a CHPID logically offline, set its status attribute to offline by writing off to it.

Issue a command of this form:

```
# echo <value> > /sys/devices/css0/chp0.<CHPID>/status
```

where:

<CHPID> is a two digit hexadecimal CHPID.

<value> is either on or off.

Examples

- To set a CHPID 0x4a logically offline issue:

```
# echo off > /sys/devices/css0/chp0.4a/status
```

- To read the status attribute to confirm that the CHPID is logically offline issue:

```
# cat /sys/devices/css0/chp0.4a/status
offline
```

- To set the same CHPID logically online issue:

```
# echo on > /sys/devices/css0/chp0.4a/status
```
To read the status attribute to confirm that the CHPID is logically online issue:

```
# cat /sys/devices/css0/chp0.4a/status
online
```

## Configuring a CHPID on LPAR

For Linux in LPAR mode, directories that represent CHPIDs contain a configure attribute that you can use to query and change the configuration state of I/O channel-paths.

### About this task

The following configuration changes are supported:

- From standby to configured (“configure”)
- From configured to standby (“deconfigure”)

### Procedure

To configure a CHPID, set its configure attribute by writing the value 1 to it. To deconfigure a CHPID, set its configure attribute by writing 0 to it.

Issue a command of this form:

```
# echo <value> > /sys/devices/css0/chp0.<CHPID>/configure
```

where:

- `<CHPID>` is a two digit hexadecimal CHPID.
- `<value>` is either 1 or 0.

To query and set the configure value using commands, see “chchp - Change channel path status” on page 492 and “lschp - List channel paths” on page 562.

### Examples

- To set a channel path with the ID 0x40 to standby issue:

```
# echo 0 > /sys/devices/css0/chp0.40/configure
```

This operation is equivalent to performing a Configure Channel Path Off operation on the hardware management console.

- To read the configure attribute to confirm that the channel path has been set to standby issue:

```
# cat /sys/devices/css0/chp0.40/configure
0
```

- To set the same CHPID to configured issue:

```
# echo 1 > /sys/devices/css0/chp0.40/configure
```
This operation is equivalent to performing a Configure Channel Path On operation on the hardware management console.

- To read the status attribute to confirm that the CHPID has been set to configured issue:

```
# cat /sys/devices/css0/chp0.40/configure
```

**Finding the physical channel associated with a CHPID**

Use the mapping of physical channel IDs (PCHID) to CHPIIDs to find the hardware from the CHPID number or the CHPID numbers from the PCHID.

**About this task**

A CHPID is associated with either a physical port or with an internal connection defined inside the mainframe, such as HiperSockets. See Figure 5. You can determine the PCHID or internal channel ID number associated with a CHPID number.

![Diagram showing relationships between CHPIIDs, PCHIDs, and internal channel ID numbers.](image)

*Figure 5. Relationships between CHPIIDs, PCHIDs, and internal channel ID numbers.*

Knowing the PCHID number can be useful in the following situations:

- When Linux indicates that a CHPID is in an error state, you can use the PCHID number to identify the associated hardware.
- When a hardware interface requires service action, the PCHID mapping can be used to determine which CHPIIDs and I/O devices will be affected.

The internal channel ID number can be useful to determine which CHPIIDs are connected to the same communication path, such as a HiperSockets link.

**Procedure**

To find the physical channel ID corresponding to a CHPID, either:

- Display the mapping of all CHPIIDs to PCHIDs. Issue the `lschp` command:

```
# lschp
```
Find the channel-ID related files for the CHPID. These sysfs files are located under /sys/devices/css0/chp0.<num>, where <num> is the two-digit, lower-case, hexadecimal CHPID number. There are two attribute files:

**chid**  
The channel ID number.

**chid_external**  
A flag indicating whether this CHPID is associated with an internal channel ID (value 0) or a physical channel ID (value 1).

The sysfs attribute files are not created when no channel ID information is available to Linux. For Linux on z/VM, the availability of this information depends on the z/VM version and configuration. For Linux in LPAR mode, this information is always available.

**Example**

The `lschp` command shows channel ID information in a column labeled PCHID. Internal channel IDs are enclosed in brackets. If no channel ID information is available, the column will show "-".

```
# lschp
CHPID Vary Cfg. Type Cmg Shared PCHID
============================================
0.30 1 1 1b 2 1 0390
0.31 1 1 1b 2 1 0392
0.32 1 1 1b 2 1 0510
0.33 1 1 1b 2 1 0512
0.34 1 0 1b - - 0580
0.fc 1 1 24 3 1 (0702)
0.fd 1 1 24 3 1 (0703)
0.fe 1 1 24 3 1 (0704)
```

This example shows that CHPID 30 is associated with PCHID 0390, while CHPID fe is associated with internal channel ID 0704.

Alternatively, check the channel ID sysfs files, for example for CHPID 30:

```
# cat /sys/devices/css0/chp0.30/chid
0390
# cat /sys/devices/css0/chp0.30/chid_external
1
```

**CCW hotplug events**

A hotplug event is generated when a CCW device appears or disappears with a machine check.

The hotplug events provide the following variables:

**CU_TYPE**  
for the control unit type of the device that appeared or disappeared.

**CU_MODEL**  
for the control unit model of the device that appeared or disappeared.

**DEV_TYPE**  
for the type of the device that appeared or disappeared.

**DEV_MODEL**  
for the model of the device that appeared or disappeared.
MODALIAS
for the module alias of the device that appeared or disappeared. The module alias is the same value that is contained in /sys/devices/css0/
<subchannel_id>/<device_bus_id>/modalias and is of the format ccw:t<cu_type>m<cu_model> or ccw:t<cu_type>m<cu_model>dt<dev_type>dm<dev_model>

Hotplug events can be used, for example, for:
• Automatically setting devices online as they appear
• Automatically loading driver modules for which devices have appeared

PCI Express support
The Peripheral Component Interconnect Express (PCIe) device driver provides support of RDMA over Converged Ethernet (RoCE).

For more information about RoCE, see Chapter 20, “RDMA over Converged Ethernet,” on page 323.

PCIe functions are seen by Linux as devices, hence devices is used here synonymously. You can assign PCIe devices to LPARs in the IOCDS.

Linux supports UIDs as persistent identifiers for PCI functions. Provide UIDs for required PCI functions in the hardware configuration (IOCDS). The LPAR needs to be enabled for UID checking. UIDs are unique hexadecimal values in the range 1 - FFFF. For example, with a UID of 0x318, the name of the function would be: 0318:00:00.0.

Setting up the PCIe support
Configure the PCIe support through the pci= kernel parameter.

PCle devices are automatically configured during the system boot process. In contrast to most IBM Z devices, all PCIe devices that are in a configured state are automatically set online. PCIe devices that are in stand-by state are not automatically enabled.

Scanning of PCIe devices is enabled by default. To disable use of PCI devices, set the kernel command line parameter pci=off.

Attention: Other PCI kernel parameters do not apply to IBM Z and might have adverse effects on your system.

PCI kernel parameter syntax

Where:
off
disables automatic scanning of PCIe devices.

on enables automatic scanning of PCIe devices (default).

Example

The following kernel parameter enables automatic scanning of PCIe devices.

```
pci=on
```

Using PCIe hotplug

Use PCIe hotplug to change the availability of a shared PCIe device.

About this task

Only one LPAR can access a PCIe device. Other LPARs can be candidates for access. Use the HMC or SE to define which LPAR is connected and which LPARs are on the candidate list. A PCIe device that is defined, but not yet used, is shown as a PCIe slot in Linux.

On Linux, you use the `power` sysfs attribute of a PCIe slot to connect the device to the LPAR where Linux runs. While a PCIe device is connected to one LPAR, it is in the reserved state for all other LPARs that are in the candidates list. A reserved PCIe device is invisible to the operating system. The slot is removed from sysfs.

Procedure

The `power` attribute of a slot contains 0 if a PCIe device is in stand-by state, or 1 if the device is configured and usable.

1. Locate the slot for the card you want to work with. To locate the slot, read the `function_id` attribute of the PCIe device from sysfs. For example, to read the `/sys/bus/pci/devices/0000:00:00.0/function_id` issue:

   ```
   # cat /sys/bus/pci/devices/0000:00:00.0/function_id
   0x00000011
   ```

   where 00000011 is the slot. Alternatively, you can use the `lspci -v` command to find the slot.

2. Write the value that you want to the `power` attribute:
   - Write 1 to `power` to connect the PCIe device to the LPAR in which your Linux instance is running. Linux automatically scans the device, registers it, and brings it online. For example:
     ```
     echo 1 > /sys/bus/pci/slots/00000011/power
     ```
   - Write 0 to `power` to stop using the PCIe device. The device state changes to stand-by. The PCIe device is set offline automatically. For example:
     ```
     echo 0 > /sys/bus/pci/slots/00000011/power
     ```

   A PCIe device in standby is also in the standby state to all other LPARs in the candidates list. A standby PCIe device appears as a slot, but without a PCIe device.

Recovering a PCIe device

Use the `recover` sysfs attribute to recover a PCIe device.
**About this task**

A message is displayed when a PCIe device enters the error state. It is not possible to automatically relieve the PCIe device from this state.

**Procedure**

1. Find the PCIe device directory in sysfs. PCIe device directories are of the form
   `/sys/devices/pci<dev>` where `<dev>` is the device ID. For example:
   `/sys/devices/pci0000:00:00.0/`

2. Write 1 to the `recover` attribute of the PCIe device. For example:
   `# echo 1 > /sys/devices/pci0000:00/0000:00:00.0/recover`

   After a successful recovery, the PCI device is de-registered and reprobed.

**Displaying PCIe information**

To display information about PCIe devices, read the attributes of the devices in sysfs.

**About this task**

The sysfs representation of a PCIe device or slot is a directory of the form
`/sys/devices/pci<device_bus_id>/<device_bus_id>`, where `<device_bus_id>` is the bus ID of the PCIe device. This sysfs directory contains a number of attributes with information about the PCIe device.

**Table 2. Read-only attributes with PCIe device information**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>function_handle</td>
<td>Eight-character, hexadecimal PCI-function (device) handle.</td>
</tr>
<tr>
<td>function_id</td>
<td>Eight-character, hexadecimal PCI-function (device) ID. The ID identifies the PCIe device within the processor configuration.</td>
</tr>
<tr>
<td>pchid</td>
<td>Four-character, hexadecimal, physical channel ID. Specifies the slot of the PCIe adapter in the I/O drawer. Thus identifies the adapter that provides the device.</td>
</tr>
<tr>
<td>pfgid</td>
<td>Two-character, hexadecimal, physical function group ID.</td>
</tr>
<tr>
<td>pfip/segment0</td>
<td>Two-character, hexadecimal, PCI-function internal path. Provides an abstract indication of the path that is used to access the PCI function. This can be used to compare the paths used by two or more PCI functions, to give an indication of the degree of isolation between them.</td>
</tr>
<tr>
<td>/segment1</td>
<td></td>
</tr>
<tr>
<td>/segment2</td>
<td></td>
</tr>
<tr>
<td>/segment3</td>
<td></td>
</tr>
<tr>
<td>uid</td>
<td>Up to eight-character, hexadecimal, user-defined identifier.</td>
</tr>
<tr>
<td>vfn</td>
<td>Four-character, hexadecimal, virtual function number. If an adapter, identified by its PCHID, supports more than one PCI function, the VFN uniquely identifies the instance of that function within the adapter.</td>
</tr>
</tbody>
</table>

**Procedure**

Issue a command of this form to read an attribute:

```shell
# cat /sys/devices/pci<device_bus_id>/<device_bus_id>/<attribute>
```

where `<attribute>` is one of the attributes of **Table 2**
Reading statistics for a PCIe device

Use the statistics attribute file to see measurement data for a PCIe device.

About this task

All PCIe devices collect measurement data by default. You can read the data in a sysfs attribute file in the debug file system, by default mounted at /sys/kernel/debug.

You can turn data collection on and off. To switch off measurement data collecting for the current session, write "0" to the statistics attribute. To enable data collection again, write "1" to the statistics attribute.

Example

To read measurement data for a (RoCE) function named 0000:00:00.0 use:

```
# cat /sys/kernel/debug/pci/0000:00:00.0/statistics
```

The statistics attribute file might look similar to this example:

```
FMB @ 0000000078cd8000
Update interval: 4000 ms
Samples: 14373
Last update TOD: cefa44fa50006378
   Load operations: 1002780
   Store operations: 1950622
   Store block operations: 0
   Refresh operations: 0
        Received bytes: 0
        Received packets: 0
        Transmitted bytes: 0
        Transmitted packets: 0
        Allocated pages: 9104
        Mapped pages: 16633
        Unmapped pages: 2337
```
Chapter 3. Kernel and module parameters

Kernel and module parameters are used to configure the kernel and kernel modules.

Individual kernel parameters or module parameters are single keywords, or keyword-value pairs of the form keyword=<{value}> with no blank. Blanks separate consecutive parameters.

Kernel parameters and module parameters are encoded as strings of ASCII characters. For tape or the z/VM reader as a boot device, the parameters can also be encoded in EBCDIC.

Use kernel parameters to configure the base kernel and any optional kernel parts that have been compiled into the kernel image. Use module parameters to configure separate kernel modules. Do not confuse kernel and module parameters. Although a module parameter can have the same syntax as a related kernel parameter, kernel and module parameters are specified and processed differently.

Where possible, this document describes kernel parameters with the device driver or feature to which they apply. Kernel parameters that apply to the base kernel or cannot be attributed to a particular device driver or feature are described in Chapter 54, “Selected kernel parameters,” on page 651. You can also find descriptions for most of the kernel parameters in Documentation/kernel-parameters.txt in the Linux source tree.

Separate kernel modules must be loaded before they can be used. Many modules are loaded automatically by Red Hat Enterprise Linux 7.5 when they are needed. To keep the module parameters in the context of the device driver or feature module to which they apply, this document describes module parameters as part of the syntax you would use to load the module with modprobe.

To find the separate kernel modules for Red Hat Enterprise Linux 7.5, list the contents of the subdirectories of /lib/modules/<kernel-release> in the Linux file system. In the path, <kernel-release> denotes the kernel level. You can query the value for <kernel-release> with uname -r.

Specifying kernel parameters

There are different methods for passing kernel parameters to the Linux kernel.

- Including kernel parameters in a boot configuration
- Using a kernel parameter file
- Specifying kernel parameters when booting Linux

Kernel parameters that you specify when booting Linux are not persistent. To define a permanent set of kernel parameters for a Linux instance, include these parameters in the boot configuration.

Note: Parameters that you specify on the kernel parameter line might interfere with parameters that Red Hat Enterprise Linux 7.5 sets for you. Read /proc/cmdline to find out which parameters were used to start a running Linux instance.
Including kernel parameters in a boot configuration

Use the zipl tool to create Linux boot configurations for IBM mainframe systems.

Which sources of kernel parameters you can use depends on the mode in which you run zipl. See "zipl modes and syntax overview" on page 56 for details.

A boot configuration can include up to 895 characters of kernel parameters. See also "How kernel parameters from different sources are combined" on page 27.

Running zipl in configuration-file mode

In configuration-file mode, you issue the zipl command with command arguments that identify a section in a zipl configuration file.

You specify details about the boot configuration in the configuration file.

As shown in Figure 6, there are three sources of kernel parameters for zipl in configuration-file mode.

In configuration-file mode, zipl concatenates the kernel parameters in the order:
1. Parameters that are specified in the kernel parameter file
2. Parameters that are specified in the zipl configuration-file
3. Parameters that are specified on the command line

See "zipl modes and syntax overview" on page 56 for details about the zipl command modes.

Running zipl in command-line mode

In command-line mode, you specify the details about the boot configuration to be created as arguments for the zipl command.

As shown in Figure 7 on page 25, there are two sources of kernel parameters for zipl in command-line mode.
In command-line mode, `zipl` concatenates the kernel parameters in the order:

1. Parameters that are specified in the kernel parameter file
2. Parameters that are specified on the command line

See “zipl modes and syntax overview” on page 56 for details about the `zipl` command modes.

Using a kernel parameter file

For booting Linux from the z/VM reader, you can use a kernel parameter file in the reader.

See “Booting from the z/VM reader” on page 88 for more details.

Specifying kernel parameters when booting Linux

Depending on the boot device and whether you boot Linux in a z/VM guest virtual machine or in LPAR mode, you can provide kernel parameters when you start the boot process.

zipl interactive boot menu on DASD

When booting Linux with a zipl interactive boot menu on a DASD boot device, you can display the menu and specify kernel parameters as you select a boot configuration. See “Example for a DASD menu configuration on z/VM” on page 86 and “Example for a DASD menu configuration (LPAR)” on page 91 for details.

z/VM guest virtual machine with a CCW boot device

When booting Linux in a z/VM guest virtual machine from a CCW boot device, you can use the PARM parameter of the IPL command to specify kernel parameters. CCW boot devices include DASD, tape, the z/VM reader, and NSS.

For details, see the subsection of “Booting Linux in a z/VM guest virtual machine” on page 84 that applies to your boot device.

z/VM guest virtual machine with a SCSI boot device

When booting Linux in a z/VM guest virtual machine from a SCSI boot device, you can use the SET LOADDEV command with the SCPDATA option to specify kernel parameters. See “Booting from a SCSI device” on page 86 for details.

LPAR mode with a SCSI boot device

When booting Linux in LPAR mode from a SCSI boot device, you can
specify kernel parameters in the **Operating system specific load parameters** field on the HMC Load panel. See Figure 26 on page 93.

Kernel parameters as entered from a CMS or CP session are interpreted as lowercase on Linux.

**Adding kernel parameters to a boot configuration**

When booting a Linux instance, you can specify kernel parameters that are used in addition to parameters in the boot configuration.

By default, the kernel parameters you specify when booting are concatenated to the end of the kernel parameters in your boot configuration. In total, the combined kernel parameter string that is used for booting can be up to 4096 characters.

If kernel parameters are specified in a combination of methods, they are concatenated in the following order:

1. Kernel parameters that have been included in the boot configuration with zipl
2. DASD only: zipl kernel parameters that are specified with the interactive boot menu
3. Depending on where you are booting Linux:
   - z/VM: kernel parameters that are specified with the PARM parameter for CCW boot devices; kernel parameters specified as SCPDATA for SCSI boot devices
   - LPAR: kernel parameters that are specified on the HMC Load panel for SCSI boot devices

If the combined kernel parameter string contains conflicting settings, the last specification in the string overrides preceding ones. Thus, you can specify a kernel parameter when booting to override an unwanted setting in the boot configuration.

**Examples**

- If the kernel parameters in your boot configuration include `possible_cpus=8` but you specify `possible_cpus=2` when booting, Linux uses `possible_cpus=2`.
- If the kernel parameters in your boot configuration include `resume=/dev/dasda2` to specify a disk from which to resume the Linux instance when it has been suspended, you can circumvent the resume process by specifying `noresume` when booting.

**Replacing all kernel parameters in a boot configuration**

Kernel parameters that you specify when booting can completely replace the kernel parameters in your boot configuration.

To replace all kernel parameters in your boot configuration, specify the new parameter string with a leading equal sign (=).

**Note:** This feature is intended for expert users who want to test a set of parameters. By replacing all parameters, you might inadvertently omit parameters that the boot configuration requires. Furthermore, you might omit parameters other than kernel parameters that Red Hat Enterprise Linux 7.5 includes in the parameter string for use by the init process.

Read `/proc/cmdline` to find out with which parameters a running Linux instance was started (see also “Displaying the current kernel parameter line” on page 28).
How kernel parameters from different sources are combined

If kernel parameters are specified in a combination of methods, they are concatenated in a specific order.

1. Kernel parameters that have been included in the boot configuration (see “Including kernel parameters in a boot configuration” on page 24). The kernel parameters in the boot configuration cannot exceed 895 characters. If more than 895 characters are specified, the excessive characters are truncated.
2. Kernel parameters that you specify through the HMC or through z/VM interfaces (see “Specifying kernel parameters when booting Linux” on page 25). For DASD boot devices you can specify up to 64 characters (z/VM only); for SCSI boot devices you can specify up to 3452 characters.

In total, the combined kernel parameter string that is passed to the Linux kernel for booting can be up to 4096 characters.

Multiple specifications for the same parameter

For some kernel parameters, multiple instances in the kernel parameter string are treated cumulatively. For example, multiple specifications for cio_ignore= are all processed and combined.

Conflicting kernel parameters

If the kernel parameter string contains kernel parameters with mutually exclusive settings, the last specification in the string overrides preceding ones. Thus, you can specify a kernel parameter when booting to override an unwanted setting in the boot configuration.

Examples:

- If the kernel parameters in your boot configuration include possible_cpus=8 but you specify possible_cpus=2 when booting, Linux uses possible_cpus=2.
- If the kernel parameters in your boot configuration include resume=/dev/dasda2 to specify a disk from which to resume the Linux instance when it has been suspended, you can circumvent the resume process by specifying noresume when booting.

Parameters other than kernel parameters

Parameters on the kernel parameter string that the kernel does not recognize as kernel parameters are ignored by the kernel and made available to user space programs. How multiple specifications and conflicts are resolved for such parameters depends on the program that evaluates them.

Examples for kernel parameters

Typical parameters that are used for booting Red Hat Enterprise Linux 7.5 configure the console and the suspend and resume function.

conmode=<mode>, condev=<cuu>, console=<name>

to set up the Linux console. See “Console kernel parameter syntax” on page 40 for details.

resume=<partition>, noresume, no_console_suspend

to configure suspend-and-resume support (see Chapter 7, “Suspending and resuming Linux,” on page 103).
See Chapter 54, “Selected kernel parameters,” on page 651 for more examples of kernel parameters.

**Displaying the current kernel parameter line**

Read `/proc/cmdline` to find out with which kernel parameters a running Linux instance was booted.

**About this task**

Apart from kernel parameters, which are evaluated by the Linux kernel, the kernel parameter line can contain parameters that are evaluated by user space programs, for example, `modprobe`.

See also “Displaying current IPL parameters” on page 98 about displaying the parameters that were used to IPL and boot the running Linux instance.

**Procedure**

Read `/proc/cmdline`. For example:

```
# cat /proc/cmdline
vconsole.keymap=us
cio_ignore=all,lcdev
crashkernel=auto
rd.zfcp=0.1707,0x500507630513c1ae,0x402140b600000000
rd.zfcp=0.1807,0x500507630508c1ae,0x402140b600000000
vconsole.font=latarcyrheb-sun16
LANG=en_US.UTF-8
BOOT_IMAGE=0
```

**Kernel parameters for rebooting**

When rebooting, you can use the current kernel parameters or an alternative set of kernel parameters.

By default, Linux uses the current kernel parameters for rebooting. See “Rebooting from an alternative source” on page 99 about how to set up Linux to use different kernel parameters for re-IPL and the associated reboot.

**Specifying parameters for modules**

How to specify parameters for modules depends on how the module is loaded, for example, automatically, through a tool, or from the command line.

You can specify parameters for modules with the `modprobe` command or on the kernel parameter line. You can specify certain parameters for modules in a boot configuration. Avoid specifying the same parameter through multiple means.

**Specifying module parameters with modprobe**

If you load a module explicitly with a modprobe command, you can specify the module parameters as command arguments.

Module parameters that are specified as arguments to modprobe are effective until the module is unloaded only.

**Note:** Parameters that you specify as command arguments might interfere with parameters that Red Hat Enterprise Linux 7.5 sets for you.
Specifying parameters on the kernel parameter line

Parameters that the kernel does not recognize as kernel parameters are ignored by the kernel and made available to user space programs.

One of these programs is modprobe, which Red Hat Enterprise Linux 7.5 uses to load modules for you. modprobe interprets module parameters that are specified on the kernel parameter line if they are qualified with a leading module prefix and a dot.

For example, you can include a specification with dasd_mod.dasd= on the kernel parameter line. modprobe evaluates this specification as the dasd= module parameter when it loads the dasd_mod module.

Including parameters for modules in a boot configuration

Parameters for modules that are required early during the boot process must be included in the boot configuration.

About this task

Red Hat Enterprise Linux 7.5 uses an initial file system (initramfs) when booting. The initramfs does not contain device specifications. Instead, it takes parameters from dracut during the boot process. dracut obtains the parameters by parsing the kernel parameter line for parameters with an “rd.” prefix.

Anaconda writes information about devices that must be accessible during the boot process to zipl.conf for you. Examples are the device with the root file system and, if configured, the swap partition that is used to resume a suspended system.

Procedure

Follow these steps to provide parameters on a kernel command line to be evaluated, for example, by dracut:

1. With an “rd.” prefix, specify the parameters in zipl.conf. For example, to specify a DASD, use rd.dasd=. The parameters are lower case and are case-sensitive. See the dracut man page, dracut.cmdline(7), for more details about parameters with an “rd.” prefix.

2. Run zipl to include the new parameter line in your boot configuration.

Displaying information about module parameters

Loaded modules can export module parameter settings to sysfs.

The parameters for modules are available as sysfs attributes of the form:

/sys/module/<module_name>/parameters/<parameter_name>

Before you begin

You can display information about modules that fulfill these prerequisites:

- The module must be loaded.
- The module must export the parameters to sysfs.

Procedure

To find and display the parameters for a module, follow these steps:
1. Optional: Confirm that the module of interest is loaded by issuing a command of this form:

```
# lsmod | grep <module_name>
```

where `<module_name>` is the name of the module.

2. Optional: Get an overview of the parameters for the module by issuing a command of this form:

```
# modinfo <module_name>
```

3. To check if a module exports settings to sysfs, try listing the module parameters. Issue a command of the form:

```
# ls /sys/module/<module_name>/parameters
```

4. If the previous command listed parameters, you can display the value for the parameter you are interested in. Issue a command of the form:

```
# cat /sys/module/<module_name>/parameters/<parameter_name>
```

**Example**

* To list the module parameters for the ap module, issue:

```
# ls /sys/module/ap/parameters
domain  ...
```

* To display the value of the domain parameter, issue:

```
# cat /sys/module/ap/parameters/domain
1
```
Part 2. Booting and shutdown

Chapter 4. Console device drivers ........................................ 33
  Console features .................................................. 33
  What you should know about the console device drivers ............ 34
  Setting up the console device drivers ................................ 40
  Working with Linux terminals ..................................... 45

Chapter 5. Initial program loader for IBM Z - zipl ............. 55
  Usage ........................................................................ 55
  Parameter overview ................................................ 72
  Configuration file structure ...................................... 76

Chapter 6. Booting Linux .................................................. 81
  IPL and booting ...................................................... 81
  Control point and boot medium ................................ 82
  Menu configurations ............................................... 82
  Boot data .................................................................. 83
  Booting Linux in a z/VM guest virtual machine .................. 84
  Booting Linux in LPAR mode ..................................... 89
  Displaying current IPL parameters ................................ 98

Chapter 7. Suspending and resuming Linux ....................... 103
  What you should know about suspend and resume .............. 103
  Setting up Linux for suspend and resume ....................... 105
  Suspending a Linux instance ...................................... 106
  Resuming a suspended Linux instance ........................... 107

Chapter 8. Shutdown actions .......................................... 109
  The shutdown configuration in sysfs ............................... 110
  Configuring z/VM CP commands as a shutdown action .......... 112

Chapter 9. The diag288 watchdog device driver ............... 115
  What you should know about the diag288 watchdog device driver .......... 115
  Loading and configuring the diag288 watchdog device driver .... 116
  External programming interfaces ................................. 118

These device drivers and features are useful for booting and shutting down instances of Red Hat Enterprise Linux 7.5.

Newest version

You can find the newest version of this book at

www.ibm.com/support/knowledgcenter/linuxonibm/iaaf/lnz_r_distlibs.html

Restrictions

For prerequisites and restrictions see the IBM Z architecture specific information in the Red Hat Enterprise Linux 7.5 release notes at

https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux

© Copyright IBM Corp. 2000, 2018
Chapter 4. Console device drivers

The console device drivers support terminal devices for basic Linux control, for example, for booting Linux, for troubleshooting, and for displaying Linux kernel messages.

The only interface to a Linux instance in an LPAR before the boot process is completed is the Hardware Management Console (HMC), see Figure 8. After the boot process has completed, you typically use a network connection to access Linux through a user login, for example, in an ssh session. The possible connections depend on the configuration of your particular Linux instance.

**Console features**

The console device drivers support several types of terminal devices.
HMC applets
You can use two applets.

Operating System Messages
This applet provides a line-mode terminal. See Figure 9 for an example.

Integrated ASCII Console
This applet provides a full-screen mode terminal.

These HMC applets are accessed through the service-call logical processor (SCLP) console interface.

3270 terminal
This terminal can be based on physical 3270 terminal hardware or a 3270 terminal emulation.

z/VM can use the 3270 terminal as a 3270 device or perform a protocol translation and use it as a 3215 device. As a 3215 device it is a line-mode terminal for the United States code page (037).

The iucvconn program
You can use the iucvconn program from Linux on z/VM to access terminal devices on other Linux instances that run as guests of the same z/VM system.

See How to Set up a Terminal Server Environment on z/VM, SC34-2596 for information about the iucvconn program.

The console device drivers support these terminals as output devices for Linux kernel messages.

Figure 9. Linux kernel messages on the HMC Operating System Messages applet

What you should know about the console device drivers
The console concepts, naming conventions, and terminology overview help you to understand the tasks you might have to perform with console and terminal devices.
Console terminology

*Terminal* and *console* have special meanings in Linux.

**Linux terminal**
An input/output device through which users interact with Linux and Linux applications. Login programs and shells typically run on Linux terminals and provide access to the Linux system.

**Linux console**
An output-only device to which the Linux kernel can write kernel messages. Linux console devices can be associated with Linux terminal devices. Thus, console output can be displayed on a Linux terminal.

**Mainframe terminal**
Any device that gives a user access to operating systems and applications that run on a mainframe. A mainframe terminal can be a physical device such as a 3270 terminal hardware that is linked to the mainframe through a controller. It can also be a terminal emulator on a workstation that is connected through a network. For example, you access z/OS® through a mainframe terminal.

**Hardware Management Console (HMC)**
A device that gives a system programmer control over IBM Z hardware resources, for example, LPARs. The HMC is a web application on a web server that is connected to the support element (SE). The HMC can be accessed from the SE but more commonly is accessed from a workstation within a secure network.

On the mainframe, the Linux console and Linux terminals can both be connected to a mainframe terminal.

**Before you have a Linux terminal - the zipl boot menu**

Do not confuse the zipl boot menu with a Linux terminal.

Depending on your setup, a zipl boot menu might be displayed when you perform an IPL. The zipl boot menu is part of the boot loader that loads the Linux kernel and is displayed before a Linux terminal is set up. The zipl boot menu is very limited in its functions. For example, there is no way to specify uppercase letters because all input is converted to lowercase characters. For more details about booting Linux, see Chapter 6, “Booting Linux,” on page 81. For more details about the zipl boot menu, see Chapter 5, “Initial program loader for IBM Z - zipl,” on page 55.

**Device and console names**
Each terminal device driver can provide a single console device.

Table 3 lists the terminal device drivers with the corresponding device names and console names.

<table>
<thead>
<tr>
<th>Device driver</th>
<th>Device name</th>
<th>Console name</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCLP line-mode terminal device driver</td>
<td>sclp_line0</td>
<td>ttyS0</td>
</tr>
<tr>
<td>SCLP VT220 terminal device driver</td>
<td>ttySclp0</td>
<td>ttyS1</td>
</tr>
<tr>
<td>3215 line-mode terminal device driver</td>
<td>ttyS0</td>
<td>ttyS0</td>
</tr>
</tbody>
</table>
Table 3. Device and console names (continued)

<table>
<thead>
<tr>
<th>Device driver</th>
<th>Device name</th>
<th>Console name</th>
</tr>
</thead>
<tbody>
<tr>
<td>3270 terminal device driver</td>
<td>3270/tty1 to 3270/tty&lt;N&gt;</td>
<td>tty3270</td>
</tr>
<tr>
<td>z/VM IUCV HVC device driver</td>
<td>hvc0 to hvc7</td>
<td>hvc0</td>
</tr>
</tbody>
</table>

As shown in Table 3 on page 35, the console with name ttyS0 can be provided either by the SCLP console device driver or by the 3215 line-mode terminal device driver. The system environment and settings determine which device driver provides ttyS0. For details, see the information about the conmode kernel parameter in “Console kernel parameter syntax” on page 40.

Of the terminal devices that are provided by the z/VM IUCV HVC device driver only hvc0 is associated with a console.

Of the 3270/tty<N> terminal devices only 3270/tty1 is associated with a console.

**Device nodes**

Applications, for example, login programs, access terminal devices by device nodes.

Table 4 shows all device nodes that udev can create for Linux instances in LPAR mode or as z/VM guests. Which of these device nodes are created depends on the conmode= and hvc_iucv= kernel parameters (see “Console kernel parameter syntax” on page 40).

Table 4. Device nodes created by udev

<table>
<thead>
<tr>
<th>Device driver</th>
<th>On LPAR</th>
<th>On z/VM</th>
<th>Major</th>
<th>Minor</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCLP line-mode terminal</td>
<td>/dev/sclp_line0</td>
<td>/dev/sclp_line0</td>
<td>4</td>
<td>64</td>
</tr>
<tr>
<td>SCLP VT220 terminal device</td>
<td>/dev/ttysclp0</td>
<td>/dev/ttysclp0</td>
<td>4</td>
<td>65</td>
</tr>
<tr>
<td>3215 line-mode terminal</td>
<td>n/a</td>
<td>/dev/ttyS0</td>
<td>4</td>
<td>64</td>
</tr>
<tr>
<td>3270 terminal device driver</td>
<td>/dev/3270/tty1</td>
<td>/dev/3270/tty1</td>
<td>227</td>
<td>1 - &lt;N&gt;</td>
</tr>
<tr>
<td>z/VM IUCV HVC device driver</td>
<td>n/a</td>
<td>/dev/hvc0</td>
<td>229</td>
<td>0 - 7</td>
</tr>
</tbody>
</table>

**Terminal modes**

The Linux terminals that are provided by the console device drivers include line-mode terminals, block-mode terminals, and full-screen mode terminals.

On a full-screen mode terminal, pressing any key immediately results in data being sent to the terminal. Also, terminal output can be positioned anywhere on the screen. This feature facilitates advanced interactive capability for terminal-based applications like the vi editor.
On a line-mode terminal, the user first types a full line, and then presses Enter to indicate that the line is complete. The device driver then issues a read to get the completed line, adds a new line, and hands over the input to the generic TTY routines.

The terminal that is provided by the 3270 terminal device driver is a traditional IBM mainframe block-mode terminal. Block-mode terminals provide full-screen output support and users can type input in predefined fields on the screen. Other than on typical full-screen mode terminals, no input is passed on until the user presses Enter. The terminal that is provided by the 3270 terminal device driver provides limited support for full-screen applications. For example, the ned editor is supported, but not vi.

Table 5 summarizes when to expect which terminal mode.

<table>
<thead>
<tr>
<th>Accessed through</th>
<th>Environment</th>
<th>Device driver</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating System Messages applet on the HMC</td>
<td>LPAR</td>
<td>SCLP line-mode terminal device driver</td>
<td>Line mode</td>
</tr>
<tr>
<td>z/VM emulation of the HMC Operating System Messages applet</td>
<td>z/VM</td>
<td>SCLP line-mode terminal device driver</td>
<td>Line mode</td>
</tr>
<tr>
<td>Integrated ASCII Console applet on the HMC</td>
<td>z/VM or LPAR</td>
<td>SCLP VT220 terminal device driver</td>
<td>Full-screen mode</td>
</tr>
<tr>
<td>3270 terminal hardware or emulation</td>
<td>z/VM with CONMODE=3215</td>
<td>3215 line-mode terminal device driver</td>
<td>Line mode</td>
</tr>
<tr>
<td>3270 terminal hardware or emulation</td>
<td>z/VM with CONMODE=3270</td>
<td>3270 terminal device driver</td>
<td>Block mode</td>
</tr>
<tr>
<td>iucvconn program</td>
<td>z/VM</td>
<td>z/VM IUCV HVC device driver</td>
<td>Full-screen mode</td>
</tr>
</tbody>
</table>

The 3270 terminal device driver provides three different views. See “Switching the views of the 3270 terminal device driver” on page 47 for details.

**How console devices are accessed**

How you can access console devices depends on your environment.

The diagrams in the following sections omit device drivers that are not relevant for the particular access scenario.

**Using the HMC for Linux in an LPAR**

You can use two applets on the HMC to access terminal devices on Linux instances that run directly in an LPAR.

Figure 10 on page 38 shows the possible terminal devices for Linux instances that run directly in an LPAR.
The **Operating System Messages** applet accesses the device that is provided by the SCLP line-mode terminal device driver. The **Integrated ASCII console** applet accesses the device that is provided by the SCLP VT220 terminal device driver.

### Using the HMC for Linux on z/VM

You can use the HMC **Integrated ASCII Console** applet to access terminal devices on Linux instances that run as z/VM guests.

While the ASCII system console is attached to the z/VM guest virtual machine where the Linux instance runs, you can access the ttyS1 terminal device from the HMC **Integrated ASCII Console** applet (see Figure 11).

### Using a 3270 terminal emulation

For Linux on z/VM, you can use 3270 terminal emulation to access a console device.

Figure 12 on page 39 illustrates how z/VM can handle the 3270 communication.
Note: Figure 12 shows two console devices with the name ttyS0. Only one of these devices can be present at any one time.

CONMODE=3215

translates between the 3270 protocol and the 3215 protocol and connects the 3270 terminal emulation to the 3215 line-mode terminal device driver in the Linux kernel.

CONMODE=3270

connects the 3270 terminal emulation to the 3270 terminal device driver in the Linux kernel.

VINPUT

is a z/VM CP command that directs input to the ttyS0 device provided by the SCLP line-mode terminal device driver. In a default z/VM environment, ttyS0 is provided by the 3215 line-mode terminal device driver. You can use the conmode kernel parameter to make the SCLP line-mode terminal device driver provide ttyS0 (see "Console kernel parameter syntax" on page 40).

The terminal device drivers continue to support 3270 terminal hardware, which, if available at your installation, can be used instead of a 3270 terminal emulation.

Using iucvconn on Linux on z/VM

On Linux on z/VM, you can access the terminal devices that are provided by the z/VM IUCV Hypervisor Console (HVC) device driver.
As illustrated in [Figure 13 on page 39], you access the devices with the iucvconn program from another Linux instance. Both Linux instances are guests of the same z/VM system. IUCV provides the communication between the two Linux instances. With this setup, you can access terminal devices on Linux instances with no external network connection.

**Note:** Of the terminal devices that are provided by the z/VM IUCV HVC device driver only hvc0 can be activated to receive Linux kernel messages.

### Setting up the console device drivers

You configure the console device drivers through kernel parameters. You also might have to enable user logins on terminals and ensure that the TERM environment variable has a suitable value.

#### Console kernel parameter syntax

Use the console kernel parameters to configure the console device drivers, line-mode terminals, and HVC terminal devices.

The `sclp_con_pages=` and `sclp_con_drop=` parameters apply only to the SCLP line-mode terminal device driver and the VT220 terminal device driver.

The `hvc_iucv=` and `hvc_iucv_allow=` kernel parameters apply only to terminal devices that are provided by the z/VM IUCV HVC device driver.

#### Console kernel parameter syntax

```
conmode= clp

console=<console_name>
```

```
sclp_con_drop=0
sclp_con_drop=1
sclp_con_pages=6
sclp_con_pages=<n>
```

```
hvc_iucv=1
hvc_iucv=<number_of_devices>
```

```
hvc_iucv_allow=<z/VM user ID>
```

**Note:** If you specify both the `conmode=` and the `console=` parameter, specify them in the sequence that is shown, `conmode=` first.

where:
conmode
specifies which one of the line-mode or block-mode terminal devices is present and provided by which device driver.

A Linux kernel might include multiple console device drivers that can provide a line-mode terminal:

- SCLP line-mode terminal device driver
- 3215 line-mode terminal device driver
- 3270 terminal device driver

On a running Linux instance, only one of these device drivers can provide a device. Table 6 shows how the device driver that is used by default depends on the environment.

Table 6. Default device driver for the line-mode terminal device

<table>
<thead>
<tr>
<th>Mode</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPAR</td>
<td>SCLP line-mode terminal device driver</td>
</tr>
<tr>
<td>z/VM</td>
<td>3215 line-mode terminal device driver or 3270 terminal device driver, depending on the z/VM guest's console settings (the CONMODE field in the output of #CP QUERY TERMINAL).</td>
</tr>
</tbody>
</table>

If the device driver you specify with the conmode= kernel parameter contradicts the CONMODE z/VM setting, z/VM is reconfigured to match the specification for the kernel parameter.

You can use the conmode parameter to override the default.

sclp or hwc
 specifies the SCLP line-mode terminal device driver.

You need this specification if you want to use the z/VM CP VINPUT command ("Using a z/VM emulation of the HMC Operating System Messages applet" on page 51).

3270
 specifies the 3270 device driver.

3215
 specifies the 3215 device driver.

console=<console_name>
specifies the console devices to be activated to receive Linux kernel messages. If present, ttyS0 is always activated to receive Linux kernel messages and, by default, it is also the preferred console.

The preferred console is used as an initial terminal device, beginning at the stage of the boot process when the initialization procedures run. Messages that are issued by programs that are run at this stage are therefore only displayed on the preferred console. Multiple terminal devices can be activated to receive Linux kernel messages but only one of the activated terminal devices can be the preferred console.

If you specify conmode=3270, there is no console with the name ttyS0.

If you want console devices other than ttyS0 to be activated to receive Linux kernel messages, specify a console statement for each of these other devices. The last console statement designates the preferred console.
If you specify one or more console parameters and you want to keep ttyS0 as the preferred console, add a console parameter for ttyS0 as the last console parameter. Otherwise, you do not need a console parameter for ttyS0.

<console_name> is the console name that is associated with the terminal device to be activated to receive Linux kernel messages. Of the terminal devices that are provided by the z/VM IUCV HVC device driver only hvc0 can be activated. Specify the console names as shown in Table 3 on page 35.

**sclp_con_drop**

governs the behavior of the SCLP line-mode terminal device driver and VT220 terminal device driver if either of them runs out of output buffer pages. The trade-off is between slowing down Linux and losing console output. Possible values are 0 (default) and 1.

0 assures complete console output by pausing until used output buffer pages are written to an output device and can be reused without loss.

1 avoids system pauses by overwriting used output buffer pages, even if the content was never written to an output device.

You can use the sclp_con_pages= parameter to set the number of output buffers.

**sclp_con_pages=<n>**

specifies the number of 4-KB memory pages to be used as the output buffer for the SCLP line-mode and VT220 terminal. Depending on the line length, each output buffer can hold multiple lines. Use many buffer pages for a kernel with frequent phases of producing console output faster than it can be written to the output device.

Depending on the setting for the sclp_con_drop=, running out of pages can slow down Linux or cause it to lose console output.

The value is a positive integer. The default is 6.

**hvc_iucv=<number_of_devices>**

specifies the number of terminal devices that are provided by the z/VM IUCV HVC device driver. <number_of_devices> is an integer in the range 0 - 8. Specify 0 to switch off the z/VM IUCV HVC device driver.

**hvc_iucv_allow=<z/VM user ID>,<z/VM user ID>, ...**

specifies an initial list of z/VM guest virtual machines that are allowed to connect to HVC terminal devices. If this parameter is omitted, any z/VM guest virtual machine that is authorized to establish the required IUCV connection is also allowed to connect. On the running system, you can change this list with the chiucvallow command. See How to Set up a Terminal Server Environment on z/VM, SC34-2596 for more information.

**Examples**

- To activate ttyS1 in addition to ttyS0, and to use ttyS1 as the preferred console, add the following specification to the kernel command line:

  ```
  console=ttyS1
  ```

- To activate ttyS1 in addition to ttyS0, and to keep ttyS0 as the preferred console, add the following specification to the kernel command line:

  ```
  console=ttyS1 console=ttyS0
  ```

- To use an emulated HMC Operating System Messages applet in a z/VM environment specify:

  ```
  conmode=sclp
  ```
To activate hvc0 in addition to ttyS0, use hvc0 as the preferred console, configure the z/VM IUCV HVC device driver to provide four devices, and limit the z/VM guest virtual machines that can connect to HVC terminal devices to lxtserv1 and lxtserv2, add the following specification to the kernel command line:

```console=hvc0  hvc_iucv=4  hvc_iucv_allow=lxtserv1,lxtserv2```

The following specification selects the SCLP line-mode terminal and configures 32 4-KB pages (128 KB) for the output buffer. If buffer pages run out, the SCLP line-mode terminal device driver does not wait for pages to be written to an output device. Instead of pausing, it reuses output buffer pages at the expense of losing content.

```console=sclp  sclp_con_pages=32  sclp_con_drop=1```

### Setting up a z/VM guest virtual machine for iucvconn

Because the iucvconn program uses z/VM IUCV to access Linux, you must set up your z/VM guest virtual machine for IUCV.

See “Setting up your z/VM guest virtual machine for IUCV” on page 320 for details about setting up the z/VM guest virtual machine.

For information about accessing Linux through the iucvttty program rather than through the z/VM IUCV HVC device driver, see How to Set up a Terminal Server Environment on z/VM, SC34-2596 or the man pages for the iucvttty and iucvconn commands.

### Setting up a line-mode terminal

The line-mode terminals are primarily intended for booting Linux.

The preferred user access to a running instance of Red Hat Enterprise Linux 7.5 is through a user login that runs, for example, in an ssh session. See “Terminal modes” on page 36 for information about the available line-mode terminals.

**Tip:** If the terminal does not provide the expected output, ensure that dumb is assigned to the TERM environment variable. For example, enter the following command on the bash shell:

```# export TERM=dumb```

### Setting up a full-screen mode terminal

The full-screen terminal can be used for full-screen text editors, such as vi, and terminal-based full-screen system administration tools.

See “Terminal modes” on page 36 for information about the available full-screen mode terminals.

**Tip:** If the terminal does not provide the expected output, ensure that linux is assigned to the TERM environment variable. For example, enter the following command on the bash shell:

```# export TERM=linux```
Setting up a terminal provided by the 3270 terminal device driver

The terminal that is provided by the 3270 terminal device driver is not a line-mode terminal, but it is also not a typical full-screen mode terminal.

The terminal provides limited support for full-screen applications. For example, the ned editor is supported, but not vi.

**Tip:** If the terminal does not provide the expected output, ensure that `linux` is assigned to the `TERM` environment variable. For example, enter the following command on the bash shell:

```
# export TERM=linux
```

Enabling user logins

Instantiate getty services for terminals to allow users access.

For the default terminals, including hvc0, systemd starts a getty. For other terminals, such as hvc1 to hvc7, you must create a symbolic link. (For the default terminals, see “Device nodes” on page 36.)

**Procedure**

To create a symbolic link and start a getty on terminal hvc1:

1. Create a new terminal instance, hvc1, for a serial getty. For example, issue:

```
# systemctl enable serial-getty@hvc1.service
```

2. Optional: Start the new instance with systemctl. For example:

```
# systemctl start serial-getty@hvc1.service
```

The terminal instance starts automatically at system start when the `getty.target` is processed.

Preventing respawns for non-operational HVC terminals

If you enable user logins on a terminal that is not available or not operational, systemd keeps respawning the getty service.

**About this task**

If user logins are enabled on unavailable HVC terminals hvc1 to hvc7, systemd might keep respawning the getty program. To be free to change the conditions that affect the availability of these terminals, use the `tyr` service to enable user logins for them. HVC terminals are operational only in a z/VM environment, and they depend on the `hvc_iucv=kernel` parameter (see “Console kernel parameter syntax” on page 40).

Any other unavailable terminals with enabled user login, including hvc0, do not cause problems with systemd.
Procedure

Perform these steps to use a ttyrun service for enabling user logins on a terminal:

1. Enable the ttyrun service by issuing a command of this form:

```bash
# systemctl enable ttyrun-getty@hvc<n>.service
```

where hvc<n> specifies one of the terminals hvc1 to hvc7.

2. Optional: Start the new service by issuing a command of this form:

```bash
# systemctl start ttyrun-getty@hvc<n>.service
```

Results

At the next system start, systemd starts the ttyrun service for hvc<n>. The ttyrun service starts a getty only if this terminal is available.

Example

For hvc1, issue:

```bash
# systemctl enable ttyrun-getty@hvc1.service
# systemctl start ttyrun-getty@hvc1.service
```

Setting up the code page for an x3270 emulation on Linux

For accessing z/VM from Linux through the x3270 terminal emulation, you must add a number of settings to the .Xdefaults file to get the correct code translation.

Add these settings:

```conf
! X3270 keymap and charset settings for Linux
x3270.charset: us-intl
x3270.keymap: circumfix
x3270.keymap.circumfix: :<key>asciicircum: Key("\^")
```

Working with Linux terminals

You might have to work with different types of Linux terminals, and use special functions on these terminals.

- “Using the terminal applets on the HMC”
- “Accessing terminal devices over z/VM IUCV” on page 46
- “Switching the views of the 3270 terminal device driver” on page 47
- “Setting a CCW terminal device online or offline” on page 48
- “Entering control and special characters on line-mode terminals” on page 49
- “Using the magic sysrequest feature” on page 49
- “Using a z/VM emulation of the HMC Operating System Messages applet” on page 51
- “Using a 3270 terminal in 3215 mode” on page 53

Using the terminal applets on the HMC

You should be aware of some aspects of the line-mode and the full-screen mode terminal when working with the corresponding applets on the HMC.
The following statements apply to both the line-mode terminal and the full-screen mode terminal on the HMC:
- On an HMC, you can open each applet only once.
- Within an LPAR, there can be only one active terminal session for each applet, even if multiple HMCs are used.
- A particular Linux instance supports only one active terminal session for each applet.
- Security hint: Always end a terminal session by explicitly logging off (for example, type “exit” and press Enter). Simply closing the applet leaves the session active and the next user to open the applet resumes the existing session without a logon.
- Slow performance of the HMC is often due to a busy console or increased network traffic.

The following statements apply to the full-screen mode terminal only:
- Output that is written by Linux while the terminal window is closed is not displayed. Therefore, a newly opened terminal window is always blank. For most applications, like login or shell prompts, it is sufficient to press Enter to obtain a new prompt.
- The terminal window shows only 24 lines and does not provide a scroll bar. To scroll up, press Shift+PgUp; to scroll down, press Shift+PgDn.

**Accessing terminal devices over z/VM IUCV**

Use z/VM IUCV to access hypervisor console (HVC) terminal devices, which are provided by the z/VM IUCV HVC device driver.

**About this task**

For information about accessing terminal devices that are provided by the iucvtty program see *How to Set up a Terminal Server Environment on z/VM, SC34-2596*.

You access HVC terminal devices from a Linux instance where the iucvconn program is installed. The Linux instance with the terminal device to be accessed and the Linux instance with the iucvconn program must both run as guests of the same z/VM system. The two guest virtual machines must be configured such that IUCV communication is permitted between them.

**Procedure**

Perform these steps to access an HVC terminal device over z/VM IUCV:
1. Open a terminal session on the Linux instance where the iucvconn program is installed.
2. Enter a command of this form:
   ```
   # iucvconn <guest_ID> <terminal_ID>
   ```
   where:
   - `<guest_ID>` specifies the z/VM guest virtual machine on which the Linux instance with the HVC terminal device to be accessed runs.
   - `<terminal_ID>` specifies an identifier for the terminal device to be accessed. HVC terminal
device names are of the form hvcn where \( n \) is an integer in the range 0-7.
The corresponding terminal IDs are lnxhvc\( n \).

**Example:** To access HVC terminal device hvc0 on a Linux instance that runs on
a z/VM guest virtual machine LXGUEST1, enter:

```
# iucvconn LXGUEST1 lnxhvc0
```

For more details and further parameters of the **iucvconn** command, see the
**iucvconn** man page or *How to Set up a Terminal Server Environment on z/VM*,
SC34-2596.

3. Press Enter to obtain a prompt.

   Output that is written by Linux while the terminal window is closed is not
displayed. Therefore, a newly opened terminal window is always blank. For
most applications, like login or shell prompts, it is sufficient to press Enter to
obtain a new prompt.

**Security hint**

Always end terminal sessions by explicitly logging off (for example, type **exit** and
press Enter). If logging off results in a new login prompt, press Control and
Underscore (Ctrl+_), then press `D` to close the login window. Simply closing the
terminal window for a hvc0 terminal device that was activated for Linux kernel
messages leaves the device active. The terminal session can then be reopened
without a login.

**Switching the views of the 3270 terminal device driver**

The 3270 terminal device driver provides three different views.

Use function key 3 (PF3) to switch between the views (see **Figure 14**).

![Figure 14. Switching views of the 3270 terminal device driver](image)

The Linux kernel messages view is available only if the terminal device is activated
for Linux kernel messages. The full-screen application view is available only if
there is an application that uses this view, for example, the ned editor.

Be aware that the 3270 terminal provides only limited full-screen support. The
full-screen application view of the 3270 terminal is not intended for applications
that require vt220 capabilities. The application itself must create the 3270 data
stream.

For the Linux kernel messages view and the terminal I/O view, you can use the
PF7 key to scroll backward and the PF8 key to scroll forward. The scroll buffers are
fixed at four pages (16 KB) for the Linux kernel messages view and five pages
(20 KB) for the terminal I/O view. When the buffer is full and more terminal data
needs to be printed, the oldest lines are removed until there is enough room. The
number of lines in the history, therefore, vary. Scrolling in the full-screen
application view depends on the application.

You cannot issue z/VM CP commands from any of the three views that are
provided by the 3270 terminal device driver. If you want to issue CP commands,
use the PA1 key to switch to the CP READ mode.

**Setting a CCW terminal device online or offline**
The 3270 terminal device driver uses CCW devices and provides them as CCW
terminal devices.

**About this task**
This section applies to Linux on z/VM. A CCW terminal device can be:

- The tty3270 terminal device that can be activated for receiving Linux kernel
  messages.
  
  If this device exists, it comes online early during the Linux boot process. In a
default z/VM environment, the device number for this device is 0009. In sysfs, it
is represented as `/sys/bus/ccw/drivers/3270/0.0.0009`. You need not set this
device online and you must not set it offline.

- CCW terminal devices through which users can log in to Linux with the CP
  DIAL command.
  
  These devices are defined with the CP DEF GRAF command. They are
represented in sysfs as `/sys/bus/ccw/drivers/3270/0.<n>.<devno>` where `<n>` is
the subchannel set ID and `<devno>` is the virtual device number. By setting these
devices online, you enable them for user logins. If you set a device offline, it can
no longer be used for user login.

See `z/VM CP Commands and Utilities Reference`, SC24-6175 for more information
about the DEF GRAF and DIAL commands.

**Procedure**
You can use the `chccwdev` command (see "chccwdev - Set CCW device attributes" on page 490) to set a CCW terminal device online or offline. Alternatively, you can
write `1` to the device's online attribute to set it online, or `0` to set it offline.

**Examples**

- To set a CCW terminal device `0.0.7b01` online, issue:

  ```
  # chccwdev -e 0.0.7b01
  ```

  Alternatively issue:

  ```
  # echo 1 > /sys/bus/ccw/drivers/3270/0.0.7b01/online
  ```

- To set a CCW terminal device `0.0.7b01` offline, issue:

  ```
  # chccwdev -d 0.0.7b01
  ```

  Alternatively issue:

  ```
  # echo 0 > /sys/bus/ccw/drivers/3270/0.0.7b01/online
  ```
### Entering control and special characters on line-mode terminals

Line-mode terminals do not have a control (Ctrl) key. Without a control key, you cannot enter control characters directly.

Also, pressing the Enter key adds a newline character to your input string. Some applications do not tolerate such trailing newline characters.

Table 7 summarizes how you can use the caret character (^) to enter some control characters and to enter strings without appended newline characters.

<table>
<thead>
<tr>
<th>For the key combination</th>
<th>Enter</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ctrl+C</td>
<td>^c</td>
<td>Cancel the process that is running in the foreground of the terminal.</td>
</tr>
<tr>
<td>Ctrl+D</td>
<td>^d</td>
<td>Generate an end of file (EOF) indication.</td>
</tr>
<tr>
<td>Ctrl+Z</td>
<td>^z</td>
<td>Stop a process.</td>
</tr>
<tr>
<td>n/a</td>
<td>^n</td>
<td>Suppresses the automatic generation of a new line. Thus, you can enter single characters; for example, the characters that are needed for yes/no answers in some utilities.</td>
</tr>
</tbody>
</table>

**Note:** For a 3215 line-mode terminal in 3215 mode, you must use United States code page (037).

### Using the magic sysrequest feature

Whether magic sysrequest functions are available and how to call them depends on your terminal.

To call the magic sysrequest functions on a line-mode terminal, enter the two characters “^-” (caret and hyphen) followed by a third character that specifies the particular function.

You can also call the magic sysrequest functions from the hvc0 terminal device if it is present and is activated to receive Linux kernel messages. To call the magic sysrequest functions from hvc0, enter the single character Ctrl+o followed by the character for the particular function.

Table 8 provides an overview of the commands for the magic sysrequest functions:

<table>
<thead>
<tr>
<th>On line-mode terminals, enter</th>
<th>On hvc0, enter</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>^-b</td>
<td>Ctrl+o b</td>
<td>Re-IPL immediately (see “Isreipl - List IPL and re-IPL settings” on page 583).</td>
</tr>
<tr>
<td>^-s</td>
<td>Ctrl+o s</td>
<td>Emergency sync all file systems.</td>
</tr>
<tr>
<td>^-u</td>
<td>Ctrl+o u</td>
<td>Emergency remount all mounted file systems read-only.</td>
</tr>
<tr>
<td>^-t</td>
<td>Ctrl+o t</td>
<td>Show task info.</td>
</tr>
<tr>
<td>^-m</td>
<td>Ctrl+o m</td>
<td>Show memory.</td>
</tr>
</tbody>
</table>
Table 8. Magic sysrequest functions (continued)

<table>
<thead>
<tr>
<th>On line-mode terminals, enter</th>
<th>On hvc0, enter</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>^- followed by a digit (0 - 9)</td>
<td>Ctrl+o</td>
<td>Set the console log level.</td>
</tr>
<tr>
<td>^-e</td>
<td>Ctrl+o e</td>
<td>Send the TERM signal to end all tasks except init.</td>
</tr>
<tr>
<td>^-i</td>
<td>Ctrl+o i</td>
<td>Send the KILL signal to end all tasks except init.</td>
</tr>
<tr>
<td>^-p</td>
<td>Ctrl+o p</td>
<td>See &quot;Obtaining debug information&quot; on page 469.</td>
</tr>
</tbody>
</table>

Note: In Table 8 on page 49, Ctrl+o means pressing 0 while holding down the control key.

Table 8 on page 49 lists the main magic sysrequest functions that are known to work on Red Hat Enterprise Linux 7.5. For a more comprehensive list of functions, see Documentation/sysrq.txt in the Linux source tree. Some of the listed functions might not work on your system.

Activating and deactivating the magic sysrequest feature

Use the sysrq procfs attribute to activate or deactivate the magic sysrequest feature.

Procedure

Issue the following command to activate the magic sysrequest function:

```
echo 1 > /proc/sys/kernel/sysrq
```

Enter the following command to deactivate the magic sysrequest feature:

```
echo 0 > /proc/sys/kernel/sysrq
```

Alternatively you can use `sysctl` to activate and deactivate the magic sysrequest feature. To check how the magic sysrequest function is set, issue:

```
# sysctl kernel.sysrq
kernel.sysrq = 1
```

In Red Hat Enterprise Linux 7.5 the magic sysrequest function is turned on by default. To turn it off using `sysctl`, issue:

```
# sysctl -w kernel.sysrq=0
```

Triggering magic sysrequest functions from procfs

If you are working from a terminal that does not support a key sequence or combination to call magic sysrequest functions, you can trigger the functions through procfs.
Procedure

Write the character for the particular function to `/proc/sysrq-trigger`. You can use this interface even if the magic syscall feature is not activated as described in “Activating and deactivating the magic syscall feature” on page 50.

Example

To set the console log level to 9, enter:

```
# echo 9 > /proc/sysrq-trigger
```

Using a z/VM emulation of the HMC Operating System Messages applet

You can use the Operating System Messages applet emulation; for example, if the 3215 terminal is not operational.

About this task

The preferred terminal devices for Linux instances that run as z/VM guests are provided by the 3215 or 3270 terminal device drivers.

The emulation requires a terminal device that is provided by the SCLP line-mode terminal device driver. To use the emulation, you must override the default device driver for z/VM environments (see “Console kernel parameter syntax” on page 40). For the emulation, you use the z/VM CP VINPUT command instead of the graphical user interface at the service element or HMC. Type any input to the operating system with a leading CP VINPUT.

The examples in the sections that follow show the input line of a 3270 terminal or terminal emulator (for example, x3270). Omit the leading #CP if you are in CP read mode. For more information about VINPUT, see z/VM CP Commands and Utilities Reference, SC24-6175.

Priority and non-priority commands

**VINPUT** commands require a VMSG (non-priority) or PVMSG (priority) specification.

Operating systems that accept this specification, process priority commands with a higher priority than non-priority commands.

The hardware console driver can accept both if supported by the hardware console within the specific machine or virtual machine.

Linux does not distinguish priority and non-priority commands.

Example

The specifications:

```
#CP VINPUT VMSG LS -L
```
and

```bash
#CP VINPUT PVMG LS -L
```

are equivalent.

**Case conversion**

All lowercase characters are converted by z/VM to uppercase. To compensate for this effect, the console device driver converts all input to lowercase.

For example, if you type `VInput VMSG echo $PATH`, the device driver gets `ECHO $PATH` and converts it into `echo $path`.

Linux and bash are case-sensitive and require some specifications with uppercase characters. To include uppercase characters in a command, use the percent sign (%) as a delimiter. The console device driver interprets characters that are enclosed by percent signs as uppercase.

**Examples**

In the following examples, the first line shows the user input. The second line shows what the device driver receives after the case conversion by CP. The third line shows the command that is processed by bash:

```
•

```bash
#cp vinput vmsg ls -l
CP VINPUT VMSG LS -L
ls -l
...
```

• The following input would result in a bash command that contains a variable $path, which is not defined in lowercase:

```
#cp vinput vmsg echo $PATH
CP VINPUT VMSG ECHO $PATH
echo $path
...
```

To obtain the correct bash command enclose the uppercase string with the conversion escape character:

```
#cp vinput vmsg echo $%PATH%
CP VINPUT VMSG ECHO $%PATH%
echo $PATH
...
```

**Using the escape character**

The quotation mark (") is the standard CP escape character. To include the escape character in a command that is passed to Linux, you must type it twice.
Example

The following command passes a string in double quotation marks to be echoed.

```bash
#cp vinput pvmsg echo "here is "$0
CP VINPUT PVMSG ECHO *HERE IS "$0
echo "here is "$0
here is -bash
```

In the example, $0 resolves to the name of the current process.

**Using the end-of-line character**

To include the end-of-line character in the command that is passed to Linux, you must specify it with a leading escape character.

If you are using the standard settings according to "Using a 3270 terminal in 3215 mode," you must specify ";# to pass # to Linux.

If you specify the end-of-line character without a leading escape character, z/VM CP interprets it as an end-of-line character that ends the VINPUT command.

**Example**

In this example a number sign is intended to mark the begin of a comment in the bash command. This character is misinterpreted as the beginning of a second command.

```bash
#cp vinput pvmsg echo "%N%umber signs start bash comments" #like this one
CP VINPUT PVMSG ECHO "%NUMBER SIGNS START BASH COMMENTS" LIKE THIS ONE
HCPCMD001E Unknown CP command: LIKE ...
```

The escape character prevents the number sign from being interpreted as an escape character.

```bash
#cp vinput pvmsg echo "%N%umber signs start bash comments" "#like this one
VINPUT PVMSG ECHO "%NUMBER SIGNS START BASH COMMENTS" #LIKE THIS ONE
echo "Number signs start bash comments" #like this one
Number signs start bash comments
```

**Simulating the Enter and Spacebar keys**

You can use the CP VINPUT command to simulate the Enter and Spacebar keys.

Simulate the Enter key by entering a blank followed by \n:

```bash
#CP VINPUT VMSG \n
```

Simulate the Spacebar key by entering two blanks followed by \n:

```bash
#CP VINPUT VMSG \n
```

**Using a 3270 terminal in 3215 mode**

The z/VM control program (CP) defines five characters as line-editing symbols. Use the CP QUERY TERMINAL command to see the current settings.
The default line-editing symbols depend on your terminal emulator. You can reassign the symbols by changing the settings of LINEND, TABCHAR, CHARDEL, LINEDEL, or ESCAPE with the CP TERMINAL command. Table 9 shows the most commonly used settings:

**Table 9. Line edit characters**

<table>
<thead>
<tr>
<th>Character</th>
<th>Symbol</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>LINEND</td>
<td>The end of line character. With this character, you can enter several logical lines at once.</td>
</tr>
<tr>
<td></td>
<td>TABCHAR</td>
<td>The logical tab character.</td>
</tr>
<tr>
<td>@</td>
<td>CHARDEL</td>
<td>The character delete symbol deletes the preceding character.</td>
</tr>
<tr>
<td>[ or ¢</td>
<td>LINEDEL</td>
<td>The line delete symbol deletes everything back to and including the previous LINEND symbol or the start of the input. “[“ is common for ASCII terminals and “¢” for EBCDIC terminals.</td>
</tr>
<tr>
<td>&quot;</td>
<td>ESCAPE</td>
<td>The escape character. With this character, you can enter a line-edit symbol as a normal character.</td>
</tr>
</tbody>
</table>

To enter a line-edit symbol, you must precede it with the escape character. In particular, to enter the escape character you must type it twice.

**Examples**

The following examples assume the settings of Table 9 with the opening bracket character (|) as the “delete line” character.

- To specify a tab character, specify:
  
  "|"

- To specify a double quotation mark character, specify:

  ""

- If you type the character string:

  #CP HALT#CP ZIPL 190[#CP IPL 10290 PARM vmpooff="MSG OP REBOOT"#CP 290"

  the actual commands that are received by CP are:

  CP HALT
  CP IPL 290 PARM vmpooff="MSG OP REBOOT"#CP 290"
Chapter 5. Initial program loader for IBM Z - zipl

Use zipl to prepare a boot device (with a Linux program loader) or to prepare a dump device.

Instead of preparing a dump device with the zipl tool you can also use the kdump infrastructure. To use kdump, no preparation with zipl is necessary. For more information about the kdump infrastructure and the dump tools that zipl installs, see Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711.

You can simulate a zipl command to test a configuration before you apply the command to an actual device (see dry-run).

zipl supports the following devices:
- Enhanced Count Key Data (ECKD) DASDs with fixed block Linux disk layout (LDL)
- ECKD DASDs with z/OS-compliant compatible disk layout (CDL)
- Fixed Block Access (FBA) DASDs
- Magnetic tape subsystems compatible with IBM3480, IBM3490, or IBM3590 (boot and dump devices only)
- SCSI with PC-BIOS disk layout

Usage

The zipl tool has base functions that can be called from the command line or in configuration-file mode. There are generic parameters and parameters that are specific to particular base functions.

zipl base functions

For each base function, there is a short and a long command-line option and, with one exception, a corresponding configuration-file option.

Table 10. zipl base functions

<table>
<thead>
<tr>
<th>Base function</th>
<th>Command line short option</th>
<th>Command line long option</th>
<th>Configuration file option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Install a boot loader</td>
<td>-i</td>
<td>--image</td>
<td>image=</td>
</tr>
<tr>
<td>Prepare a DASD or tape dump device</td>
<td>-d</td>
<td>--dumpto</td>
<td>dumpto=</td>
</tr>
<tr>
<td>Prepare a list of ECKD volumes for a multi-volume dump</td>
<td>-M</td>
<td>--mvdump</td>
<td>mvdump=</td>
</tr>
</tbody>
</table>

See “Preparing a boot device” on page 59 for details.

See “Preparing a DASD or tape dump device” on page 65 for details.

See “Preparing a multi-volume dump on ECKD DASD” on page 67 for details.
Table 10. `zipl` base functions (continued)

<table>
<thead>
<tr>
<th>Base function</th>
<th>Command line short option</th>
<th>Command line long option</th>
<th>Configuration file option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepare a SCSI dump device</td>
<td>-D</td>
<td>--dumptofs</td>
<td>dumptofs=</td>
</tr>
<tr>
<td>See “Preparing a dump device on a SCSI disk” on page 68 for details.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Install a menu configuration</td>
<td>-m</td>
<td>--menu</td>
<td>(None)</td>
</tr>
<tr>
<td>See “Installing a menu configuration” on page 71 for details.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

zipl modes and syntax overview

When running `zipl`, you can either directly specify a base function with its parameters or a configuration file with specifications, or you can use the default `zipl` configuration file.

`zipl` operates in one of two modes:

Command-line mode

If a `zipl` command is issued with a base function other than installing a menu configuration (see “Installing a menu configuration” on page 71), the entire configuration must be defined by using command-line parameters. See the following base functions for how to specify command-line parameters:

- “Preparing a boot device” on page 59
- “Preparing a DASD or tape dump device” on page 65
- “Preparing a multi-volume dump on ECKD DASD” on page 67
- “Preparing a dump device on a SCSI disk” on page 68

Configuration-file mode

If a `zipl` command is issued either without a base function or to install a menu configuration, a configuration file is accessed. For more information, see “Configuration file structure” on page 76.
zipl syntax overview

-zipl  -V  --dry-run

parameters when omitting base function:

- i  i_parameters
- d  d_parameters
- M  M_parameters
- D  D_parameters
- m  m_parameters

parameters when omitting base function:

(1) -c /etc/zipl.conf
(2) -c <config_file>
(3) -P <parameters>
(4) -a

Notes:
1. You can change the default configuration file with the ZIPLCONF environment variable.
2. If no configuration is specified, zipl uses the configuration in the [defaultboot] section of the configuration file (see “Configuration file structure” on page 76).
3. In a boot configuration or with a SCSI dump configuration only.
4. In a boot configuration or a menu configuration only.

Where:

- c <config_file>
specifies the configuration file to be used.

<configuration>
specifies a single configuration section in a configuration file.

-P <parameters>
can optionally be used to provide:

kernel parameters
in a boot configuration section. See “How kernel parameters from different sources are combined” on page 61 for information about how kernel parameters specified with the -P option are combined with any kernel parameters specified in the configuration file.

SCSI system dumper parameters
in a SCSI dump configuration section. See “How SCSI system dumper parameters from different sources are combined” on page 71 for information about how parameters specified with the -P option are combined with any parameters specified in the configuration file.
If you provide multiple parameters, separate them with a blank and enclose them within single quotation marks (') or double quotation marks (").

- **a** in a boot configuration section, adds kernel image, kernel parameter file, and initial RAM disk to the bootmap file. Use this option when these files are spread across multiple disks to ensure that they are available at IPL time. Specifying this option significantly increases the size of the bootmap file that is created in the target directory.

- **n** suppresses confirmation prompts that require operator responses to allow unattended processing (for example, for processing DASD or tape dump configuration sections).

- **V** provides verbose command output.

- **--dry-run**
  simulates a **zipl** command. Use this option to test a configuration without overwriting data on your device.
  During simulation, **zipl** performs all command processing and issues error messages where appropriate. Data is temporarily written to the target directory and is cleared up when the command simulation is completed.

- **v** displays version information.

- **h** displays help information.

The basic functions and their parameters are described in detail in the following sections.

See "Parameter overview" on page 72 for a summary of the short and long command-line options and their configuration file equivalents.

**Examples**

- To process the default configuration in the default configuration file (**/etc/zipl.conf**, unless specified otherwise with the environment variable **ZIPLCONF**) issue:

  # zipl

- To process the default configuration in a configuration file **/etc/myxmp.conf** issue:

  # zipl -c /etc/myxmp.conf

- To process a configuration [myconf] in the default configuration file issue:

  # zipl myconf

- To process a configuration [myconf] in a configuration file **/etc/myxmp.conf** issue:

  # zipl -c /etc/myxmp.conf myconf

- To simulate processing a configuration [myconf] in a configuration file **/etc/myxmp.conf** issue:

  # zipl --dry-run -c /etc/myxmp.conf myconf
Preparing a boot device

Use `zipl` with the `-i (--image)` command-line option or with the `image=` configuration-file option to prepare a boot device.

zipl command line syntax for preparing a boot device

```
zipl -i <image> ,<image_addr>
-t <directory> (1)
-R <tape_node>
-r <ramdisk> ,<initrd_addr>
-p <parmfile> ,<parm_addr>
-P <parameters> -k auto
```

Notes:
1 Additional parameters that are used only if `-t` specifies a logical device as a target. See “Using base device parameters” on page 63.

To prepare a device as a boot device, you must specify:

**The location `<image>`**

of the Linux kernel image on the file system.

**A target `<directory>` or `<tape_node>`**

`zipl` installs the boot loader code on the device that contains the specified directory `<directory>` or to the specified tape device `<tape_node>`.

Optionally, you can also specify:

**A kernel image address `<image_addr>`**

to which the kernel image is loaded at IPL time. The default address is 0x10000.

**The RAM disk location `<ramdisk>`**

of an initial RAM disk image (initrd) on the file system.
A RAM disk image address `<initrd_addr>`
to which the RAM disk image is loaded at IPL time. If you do not specify
this parameter, `zipl` investigates the location of other components and
calculates a suitable address for you.

Kernel parameters
to be used at IPL time. If you provide multiple parameters, separate them
with a blank and enclose them within single quotation marks (') or double
quotation marks (").

You can specify parameters `<parameters>` directly on the command line.
Instead or in addition, you can specify a location `<parmfile>` of a kernel
parameter file on the file system. See “How kernel parameters from
different sources are combined” on page 61 for a discussion of how `zipl`
combines multiple kernel parameter specifications.

A parameter address `<parm_addr>`
to which the kernel parameters are loaded at IPL time. The default address
is 0x1000.

An option `-k auto`
to install a kdump kernel that can be used as a stand-alone dump tool. You
can IPL this kernel in an LPAR or guest virtual machine to create a dump
of a previously running operating system instance that has been configured
with a reserved memory area for kdump. For Linux, this memory area is
reserved with the `crashkernel=` kernel parameter.

Note: For SCSI disks, the accumulated size of the kernel and ramdisk must
not exceed 16 MB.

An option `-a`
to add the kernel image, kernel parameter file, and initial RAM disk to the
bootmap file. Use this option when these files are spread across multiple
disks to ensure that they are available at IPL time. This option is available
on the command line only. Specifying this option significantly increases the
size of the bootmap file that is created in the target directory.

See “Parameter overview” on page 72 for a summary of the parameters. This
summary includes the long options that you can use on the command line.

Figure 15 on page 61 summarizes how you can specify a boot configuration within
a configuration file section. Required specifications are shown in bold. See
“Configuration file structure” on page 76 for a more comprehensive discussion of
the configuration file.
The following command identifies the location of the kernel image as /boot/mnt/image-2, identifies the location of an initial RAM disk as /boot/mnt/initrd, specifies a kernel parameter file /boot/mnt/parmf-2, and writes the required boot loader code to /boot. At IPL time, the initial RAM disk is to be loaded to address 0x900000, rather than an address that is calculated by zipl. Kernel image, initial RAM disk, and the kernel parameter file are to be copied to the bootmap file on the target directory /boot rather than being referenced.

An equivalent section in a configuration file might look like this example:

```
[boot2]
image=/boot/mnt/image-2
ramdisk=/boot/mnt/initrd,0x900000
parmf=/boot/mnt/parmf-2
parameters=/boot/mnt/parmf-2
# Next line for devices other than tape only
target=/directory
# Next line for tape devices only
tape=<tape_node>
# Next line for stand-alone kdump only
tape=auto
```

Figure 15. zipl syntax for preparing a boot device - configuration file mode

**Example**

The following command identifies the location of the kernel image as /boot/mnt/image-2, identifies the location of an initial RAM disk as /boot/mnt/initrd, specifies a kernel parameter file /boot/mnt/parmf-2, and writes the required boot loader code to /boot. At IPL time, the initial RAM disk is to be loaded to address 0x900000, rather than an address that is calculated by zipl. Kernel image, initial RAM disk, and the kernel parameter file are to be copied to the bootmap file on the target directory /boot rather than being referenced.

```
# zipl -i /boot/mnt/image-2 -r /boot/mnt/initrd,0x900000 -p /boot/mnt/parmf-2 -t /boot -a
```

An equivalent section in a configuration file might look like this example:

```
[boot2]
image=/boot/mnt/image-2
ramdisk=/boot/mnt/initrd,0x900000
parmf=/boot/mnt/parmf-2
parameters=/boot/mnt/parmf-2
target=/directory
```

There is no configuration file equivalent for option -a. To use this option for a boot configuration in a configuration file, it must be specified with the zipl command that processes the configuration.

If the configuration file is called /etc/myxmp.conf:

```
# zipl -c /etc/myxmp.conf boot2 -a
```

**How kernel parameters from different sources are combined**

zipl allows for multiple sources of kernel parameters when preparing boot devices.

In command-line mode, there are two possible sources of kernel parameters. The parameters are processed in the following order:

1. Parameters in the kernel parameter file (specified with the -p or --parmf option)
2. Parameters that are specified on the command line (specified with the -P or --parameters option)

In configuration file mode, there are three possible sources of kernel parameters. The parameters are processed in the following order:
1. Parameters that are specified in the kernel parameter file (specified with the `parmfile=` option)
2. Parameters that are specified in the configuration section (specified with the `parameters=` option)
3. Parameters that are specified on the command line (specified with the `-P` or `--parameters` option)

Parameters from different sources are concatenated and passed to the kernel in one string. At IPL time, the combined kernel parameter string is loaded to address 0x1000, unless an alternate address is provided.

For more information about the different sources of kernel parameters, see "Including kernel parameters in a boot configuration" on page 24.

**Preparing a logical device as a boot device**

A *logical device* is a block device that represents one or more real devices.

If your boot directory is on a logical DASD or SCSI device, *zipl* cannot detect all required information about the underlying real device or devices and needs extra input.

Logical devices can be two DASDs combined into a logical mirror volume. Another examples are a linear mapping of a partition to a real device or a more complex mapping hierarchy. Logical devices are controlled by a device mapper.

Blocks on the logical device must map to blocks on the underlying real device or devices linearly. If two blocks on the logical device are adjacent, they must also be adjacent on the underlying real devices. This requirement excludes mappings such as striping.

You always boot from a real device. *zipl* must be able to write to that device, starting at block 0. In a logical device setup, starting at the top of the mapping hierarchy, the first block device that grants access to block 0 (and subsequent blocks) is the *base device*, see [Figure 16](#).

![Figure 16. Definition of base device](#)

A base device can have the following mappings:

- A mapping to a part of a real device that contains block 0
A mapping to one complete real device
- A mapping to multiple real devices.

For a mapping to multiple real devices all the real devices must share the device characteristics and contain the same data (for example, a mirror setup). The mapping can also be to parts of the devices if these parts contain block 0. The mapping must not combine multiple devices into one large device.

The `zipl` command needs the device node of the base device and information about the physical characteristics of the underlying real devices. For most logical boot devices, a helper script automatically provides all the required information to `zipl` for you (see “Using a helper script”).

If you decide not to use the supplied helper script, or want to write your own helper script, you can use parameters to supply the base device information to `zipl`, see “Using base device parameters” and “Writing your own helper script” on page 65.

**Using a helper script**

`zipl` provides a helper script, `zipl_helper.device-mapper`, that detects the required information and provides it to `zipl` for you.

The helper script is used automatically when you run `zipl` to prepare a boot device. Specify the parameters for the kernel image, parameter file, initial RAM disk, and target as usual. See “Preparing a boot device” on page 59 for details about the parameters.

Assuming an example device for which the location of the kernel image is `/boot/image-5`, the location of an initial RAM disk as `/boot/initrd-5`, a kernel parameter file `/boot/parmf-5`, and which writes the required boot loader code to `/boot` and is a device mapper device, the command then becomes:

```
# zipl -i /boot/image-5 -r /boot/initrd-5 -p /boot/parmf-5 -t /boot
```

The corresponding configuration file section becomes:

```
[boot5]
image=/boot/image-5
ramdisk=/boot/initrd-5
paramfile=/boot/parmf-5
target=/boot
```

**Using base device parameters**

You can use parameters to supply the base device information to `zipl` directly.

The following command syntax for the base device parameters is used for logical boot devices. It extends the `zipl` syntax as shown in “Preparing a boot device” on page 59.
You must specify the following device information:

**The device node `<targetbase_node>`**

of the base device, either by using the standard device name or in form of the major and minor number, separated by a colon (:).

*Example:* The device node specification for the device might be `/dev/dm-0` and the equivalent specification with major and minor numbers might be `253:0`.

**The device type**

of the base device. The following specifications are valid:
- **LDL** for ECKD type DASD with the Linux disk layout.
- **CDL** for ECKD type DASD with the compatible disk layout.
- **FBA** for FBA type DASD.
- **SCSI** for FCP-attached SCSI disks.

**LDL and CDL only:** The disk geometry `<cylinders>,<heads>,<sectors>`

of the base device in cylinders, heads, and sectors.

**The block size `<targetblocksize>`**

in bytes per block of the base device.

**The offset `<targetoffset>`**

in blocks between the start of the physical device and the start of the topmost logical device in the mapping hierarchy.

Figure 17 shows how you can specify this information in a configuration file.

```
[<section_name>]
image=<image>,<image_addr>
ramdisk=<ramdisk>,<initrd_addr>
parmfile=<parmfile>,<parm_addr>
parameters=<parameters>
target=<directory>
targetbase=<targetbase_node>
targettype=LDL|CDL|FBA|SCSI
  # Next line for target types LDL and CDL only
  targetgeometry=<cylinders>,<heads>,<sectors>
targetblocksize=<targetblocksize>
targetoffset=<targetoffset>
```

*Figure 17.* zipl syntax for preparing a logical device as a boot device - configuration file mode
Example

The example command identifies the location of the kernel image as /boot/image-5, identifies the location of an initial RAM disk as /boot/initrd-5, specifies a kernel parameter file /boot/parmf-5, and writes the required boot loader code to /boot.

The command specifies the following information about the base device: the device node is /dev/dm-3, the device has the compatible disk layout, there are 6678 cylinders, there are 15 heads, there are 12 sectors, and the topmost logical device in the mapping hierarchy begins with an offset of 24 blocks from the start of the base device.

```
# zipl -i /boot/image-5 -r /boot/initrd-5 -p /boot/parmf-5 -t /boot --targetbase /dev/dm-3 \ 
# --targettype CDL --targetgeometry 6678,15,12 --targetblocksize=4096 --targetoffset 24
```

Note: Instead of using the continuation sign (\) at the end of the first line, you might want to specify the entire command on a single line.

An equivalent section in a configuration file might look like this example:

```
[boot5]
image=/boot/image-5
ramdisk=/boot/initrd-5
paramfile=/boot/parmf-5
target=/boot
targetbase=/dev/dm-3
targettype=CDL
targetgeometry=6678,15,12
targetblocksize=4096
targetoffset=24
```

Writing your own helper script

You can write your own helper script for device drivers that provide logical devices. The helper script must conform to a set of rules.

- The script must accept the name of the target directory as an argument. From this specification, it must determine a suitable base device. See “Using base device parameters” on page 63.
- The script must write the following base device <parameter>=<value> pairs to stdout as ASCII text. Each pair must be written on a separate line.
  - targetbase=<targetbase_node>
  - targettype=<type> where type can be LDL, CDL, FBA, or SCSI.
  - targetgeometry=<cylinders>,<heads>,<sectors> (For LDL and CDL only)
  - targetblocksize=<blocksize>
  - targetoffset=<offset>
  See “Using base device parameters” on page 63 for the meaning of the base device parameters.
- The script must be named zipl_helper.<device> where <device> is the device name as specified in /proc/devices.
- The script must be in /lib/s390-tools.

Preparing a DASD or tape dump device

Use zipl with the -d (--)dump-to option or with the dump-to= configuration-file option to prepare a DASD or tape dump device.
zipl command line syntax for preparing a DASD or tape dump device

```
zipl -d <dump_device>,<size> -n
```

To prepare a DASD or tape dump device, you must specify:

The **device node** `<dump_device>`

of the DASD partition or tape device to be prepared as a dump device. `zipl` deletes all data on the partition or tape and installs the boot loader code there.

**Note:**

- If the dump device is an ECKD disk with fixed-block layout (LDL), a dump overwrites the dump utility. You must reinstall the dump utility before you can use the device for another dump.
- If the dump device is a tape, FBA disk, or ECKD disk with the compatible disk layout (CDL), you do not need to reinstall the dump utility after every dump.

Optionally, you can also specify:

**An option** `-n`

to suppress confirmation prompts to allow unattended processing (for example, from a script). This option is available on the command line only.

**A limit** `<size>`

for the amount of memory to be dumped. The value is a decimal number that can optionally be suffixed with K for kilobytes, M for megabytes, or G for gigabytes. The value is rounded to the next megabyte boundary.

If you limit the dump size below the amount of memory that is used by the system to be dumped, the resulting dump is incomplete.

DASD or tape dump devices are not formatted with a file system so no target directory can be specified. See *Using the Dump Tools on Red Hat Enterprise Linux 7*, SC34-2711 for details about processing these dumps.

See “Parameter overview” on page 72 for a summary of the parameters. The summary includes the long options that you can use on the command line.

**Figure 18** summarizes how you can specify a DASD or tape dump configuration in a configuration file. See “Configuration file structure” on page 76 for a more comprehensive discussion of the configuration file.

```
[<section_name>]
dumpto=<dump_device>,<size>
```

**Figure 18. zipl syntax for preparing a DASD or tape dump device - configuration file mode**
Example

The following command prepares a DASD partition /dev/dasdc1 as a dump device and suppresses confirmation prompts that require an operator response:

```
# zipl -d /dev/dasdc1 -n
```

An equivalent section in a configuration file might look like this example:
```
[dumpdasd]
dumpto=/dev/dasdc1
```

There is no configuration file equivalent for option `-n`. To use this option for a DASD or tape dump configuration in a configuration file, it must be specified with the `zipl` command that processes the configuration.

If the configuration file is called `/etc/myxmp.conf`:
```
# zipl -c /etc/myxmp.conf dumpdasd -n
```

Preparing a multi-volume dump on ECKD DASD

Use `zipl` with the `-M (--mvdump)` command-line option or with the `mvdump=` configuration-file option to prepare a multi-volume dump on ECKD DASD.

```
zipl command line syntax for preparing devices for a multi-volume dump

```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
```-
If you limit the dump size below the amount of memory that is used by the system to be dumped, the resulting dump is incomplete.

DASD or tape dump devices are not formatted with a file system so no target directory can be specified. See Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711 for details about processing these dumps.

See “Parameter overview” on page 72 for a summary of the parameters. This summary includes the long options that you can use on the command line.

Figure 19 summarizes how you can specify a multi-volume DASD dump configuration in a configuration file. See “Configuration file structure” on page 76 for a more comprehensive discussion of the configuration file.

```
[<section_name>]
mvdump=<dump_device_list>,<size>
```

Figure 19. `zipl` syntax for preparing DASD devices for a multi-volume dump - configuration file mode

**Example**

The following command prepares two DASD partitions `/dev/dasdc1`, `/dev/dasdd1` for a multi-volume dump and suppresses confirmation prompts that require an operator response:

```
# zipl -M mvdump.conf -n
```

where the `mvdump.conf` file contains the two partitions that are separated by line breaks:

```
/dev/dasdc1
/dev/dasdd1
```

An equivalent section in a configuration file might look like this example:

```
[multi_volume_dump]
mvdump=mvdump.conf
```

There is no configuration file equivalent for option `-n`. To use this option for a multi-volume DASD dump configuration in a configuration file, it must be specified with the `zipl` command that processes the configuration.

If the configuration file is called `/etc/myxmp.conf`:

```
# zipl -c /etc/myxmp.conf multi_volume_dump -n
```

**Preparing a dump device on a SCSI disk**

Use `zipl` with the `-D (--dumptofs)` command-line option or with the `dumptofs=` configuration-file option to prepare a dump device on a SCSI disk.

**Before you begin**: At least one partition, the *target partition*, must be available to `zipl`. 
zipl command line syntax for preparing a SCSI dump device

```
zipl -D <dump_partition>,<size> -t <directory>

- P <parameters> - p <parmfile>
```

The target partition contains the target directory and is accessed to load the SCSI system dumper tool at IPL time. Dumps are written as files to a *dump partition*.

The dump and target partition can but need not be the same partition. Preferably, dump and target partition are two separate partitions.

The target and dump partitions must be formatted with a file system supported by the SCSI Linux system dumper tool. Unlike DASD and tape, creating a dump device on SCSI disk does not destroy the contents of the target partition. See *Using the Dump Tools on Red Hat Enterprise Linux 7*, SC34-2711 for more details.

To prepare a SCSI disk as a dump device, you must specify:

**The dump partition** `<dump_partition>`

to which the dumps are written.

**A target** `<directory>`

to which the SCSI system dumper components are written. `zipl` uses the target directory to determine the dump device (target partition).

Optionally, you can also specify:

**SCSI system dumper parameters**

You can specify parameters `<parameters>` directly on the command line. Instead or in addition, you can specify a location `<parmfile>` of a parameter file on the file system. See "How SCSI system dumper parameters from different sources are combined" on page 71 for a discussion of how multiple parameter specifications are combined.

**dump_dir=/<directory>**

Path to the directory (relative to the root of the dump partition) where the dump file is to be written. This directory is specified with a leading slash. The directory must exist when the dump is initiated.

**Example:** If the dump partition is mounted as `/dumps`, and the parameter `dump_dir=/mydumps` is defined, the dump directory would be accessed as `/dumps/mydumps`.

The default is `/` (the root directory of the partition).

**dump_compress=gzip | none**

Dump compression option. Compression can be time-consuming on slower systems with a large amount of memory.

The default is `none`. 

**dump_mode=interactive | auto**

Action that is taken if there is no room on the file system for the new dump file. interactive prompts the user to confirm that the dump with the lowest number is to be deleted. auto automatically deletes this file.

The default is interactive.

If you provide multiple parameters, separate them with a blank and enclose them within single quotation marks (') or double quotation marks (").

**A limit <size>**

for the amount of memory to be dumped. The value is a decimal number that can optionally be suffixed with K for kilobytes, M for megabytes, or G for gigabytes. The value is rounded to the next megabyte boundary.

If you limit the dump size below the amount of memory that is used by the system to be dumped, the resulting dump is incomplete.

See "Parameter overview" on page 72 for a summary of the parameters. This summary includes the long options that you can use on the command line.

Figure 20 summarizes how you can specify a SCSI dump configuration in a configuration file. Required specifications are shown in bold. See "Configuration file structure" on page 76 for a more comprehensive discussion of the configuration file.

```
[<section_name>]
dumptofs=<dump_partition>
parmfile=<parmfile>,<parm_addr>
parameters=<parameters>
target=<directory>
```

Figure 20. zipl syntax for preparing a SCSI dump device - configuration file mode

**Example**

The following command prepares a SCSI partition /dev/sda2 as a dump device and a directory /boot as the target directory. Dumps are to be written to a directory mydumps, relative to the mount point. There is to be no compression but instead the oldest dump is automatically deleted if there is not enough space for the new dump.

```
# zipl -D /dev/sda2 -P 'dumpdir=/mydumps dump_compress=none dump_mode=auto' -t /boot
```

An equivalent section in a configuration file might look like this example:

```
[dumpscsi]
dumptofs=/dev/sda2
parameters="dumpdir=/mydumps dump_compress=none dump_mode=auto"
target=/boot
```

In both the command line and configuration file examples the parameter specifications dump_compress=none dump_mode=auto could be omitted because they correspond to the defaults.
If the configuration file is called /etc/myxmp.conf, the `zipl` command that processes the configuration would be:

```
# zipl -c /etc/myxmp.conf dumpscsi
```

### How SCSI system dumper parameters from different sources are combined
`zipl` allows for multiple sources of SCSI system dumper parameters.

In command-line mode, there are two possible sources of parameters. The parameters are processed in the following order:
1. Parameters in a parameter file (specified with the `-p` or `--parmfile` option)
2. Parameters that are specified on the command line (specified with the `-P` or `--parameters` option)

In configuration file mode, there are three possible sources of parameters. The parameters are processed in the following order:
1. Parameters in a parameter file (specified with the `parmfile=` option)
2. Parameters that are specified in the configuration section (specified with the `parameters=` option)
3. Parameters that are specified on the command line (specified with the `-P` or `--parameters` option)

Parameters from different sources are concatenated and passed to the SCSI system dumper in one string. If the same parameter is specified in multiple sources, the value that is encountered last is used. At IPL time, the combined parameter string is loaded to address (0x1000).

### Installing a menu configuration
Use `zipl` with the `-m` (`--menu`) command-line option to install a menu configuration.

To prepare a menu configuration, you need a configuration file that includes at least one menu.

```
zipl syntax for installing a menu configuration
```

```
zipl -m <menu_name> -c /etc/zipl.conf -c <config_file> -a
```

**Notes:**
1. You can change the default configuration file with the ZIPLCONF environment variable.

**Where:**
-m or --menu
specifies the menu that defines the menu configuration in the configuration file.

<config_file>
specifies the configuration file where the menu configuration is defined. The default, /etc/zipl.conf, can be changed with the ZIPLCONF environment variable.

-a or --add-files
adds the kernel image file, parmfile, and initial RAM disk image to the bootstrap files in the respective target directories instead of referencing them. Use this option if the files are spread across disks to ensure that the files are available at IPL time. Specifying this option significantly increases the size of the bootstrap file that is created in the target directory.

Example

Using the sample configuration file of Figure 21 on page 79, you could install a menu configuration with:

```
# zipl -m menu1
```

Parameter overview

You might need to know all zipl options and how to specify them on the command line or in the configuration file.

<table>
<thead>
<tr>
<th>Command line short option</th>
<th>Command line long option</th>
<th>Configuration file option</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>-a</td>
<td>--add-files</td>
<td>n/a</td>
<td>Causes kernel image, kernel parameter file, and initial RAM disk to be added to the bootstrap file in the target directory rather than being referenced from this file. Use this option when these files are spread across multiple disks to ensure that they are available at IPL time. Specifying this option significantly increases the size of the bootstrap file that is created in the target directory.</td>
</tr>
<tr>
<td>-c &lt;config_file&gt;</td>
<td>--config=&lt;config_file&gt;</td>
<td>n/a</td>
<td>Specifies the configuration file. You can change the default configuration file /etc/zipl.conf with the environment variable ZIPLCONF.</td>
</tr>
<tr>
<td>&lt;configuration&gt;</td>
<td>n/a</td>
<td>n/a</td>
<td>Specifies a configuration section to be read and processed from the configuration file.</td>
</tr>
<tr>
<td>Command line short option</td>
<td>Command line long option</td>
<td>Explanation</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>--------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td><strong>-d</strong> &lt;dump_device&gt;[,&lt;size&gt;]</td>
<td><strong>--dumpeto=&lt;dump_device&gt;[,&lt;size&gt;]</strong></td>
<td>Specifies the DASD partition or tape device to which a dump is to be written after IPL. The optional size specification limits the amount of memory to be dumped. The value is a decimal number that can optionally be suffixed with K for kilobytes, M for megabytes, or G for gigabytes. The value is rounded to the next megabyte boundary. If you limit the dump size below the amount of memory that is used by the system to be dumped, the resulting dump is incomplete. If no limit is provided, all of the available physical memory is dumped. See “Preparing a DASD or tape dump device” on page 65 and Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711 for details.</td>
<td></td>
</tr>
<tr>
<td><strong>-D</strong> &lt;dump_partition&gt;[,&lt;size&gt;]</td>
<td><strong>--dumpertos=&lt;dump_partition&gt;[,&lt;size&gt;]</strong></td>
<td>Specifies the partition to which a SCSI dump file is to be written. This partition must be formatted with a file system supported by the SCSI Linux system dumper tool. The dump partition must be on the same physical SCSI disk as the target partition. It can but need not be the partition that also contains the target directory (target partition). The optional size specification limits the amount of memory to be dumped. The value is a decimal number that can optionally be suffixed with K for kilobytes, M for megabytes, or G for gigabytes. The value is rounded to the next megabyte boundary. If you limit the dump size below the amount of memory that is used by the system to be dumped, the resulting dump is incomplete. If no limit is provided, all of the available physical memory is dumped. See “Preparing a dump device on a SCSI disk” on page 68 and Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711 for details.</td>
<td></td>
</tr>
<tr>
<td><strong>-h</strong></td>
<td><strong>--help</strong></td>
<td>Displays help information.</td>
<td></td>
</tr>
<tr>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>-i</strong> &lt;image&gt;[,&lt;image_addr&gt;]</td>
<td><strong>--image=&lt;image&gt;[,&lt;image_addr&gt;]</strong></td>
<td>Specifies the location of the Linux kernel image on the file system and, optionally, in memory after IPL. The default memory address is 0x10000. See “Preparing a boot device” on page 59 for details.</td>
<td></td>
</tr>
<tr>
<td><strong>-k</strong> auto</td>
<td><strong>--kdump=auto</strong></td>
<td>Installs a kdump kernel that can be used as a stand-alone dump tool. You can IPL this kernel in an LPAR or guest virtual machine to create a dump of a previously running operating system instance that has been configured with a reserved memory area for kdump. For Linux, this memory area is reserved with the crashkernel= kernel parameter. See “Preparing a boot device” on page 59 for details.</td>
<td></td>
</tr>
<tr>
<td>Command line short option</td>
<td>Command line long option</td>
<td>Configuration file option</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------------------------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>-m &lt;menu_name&gt;</td>
<td>--menu=&lt;menu_name&gt;</td>
<td>n/a</td>
<td>Specifies the name of the menu that defines a menu configuration in the configuration file (see “Menu configurations” on page 77).</td>
</tr>
<tr>
<td>-M &lt;dump_device_list&gt;[,&lt;size&gt;]</td>
<td>--mvdump=&lt;dump_device_list&gt;[,&lt;size&gt;]</td>
<td>n/a</td>
<td>Specifies a file with a list of DASD partitions to which a dump is to be written after IPL.</td>
</tr>
<tr>
<td></td>
<td>mvdump=&lt;dump_device_list&gt;[,&lt;size&gt;]</td>
<td></td>
<td>The optional size specification limits the amount of memory to be dumped. The value is a decimal number that can optionally be suffixed with K for kilobytes, M for megabytes, or G for gigabytes. The value is rounded to the next megabyte boundary. If you limit the dump size below the amount of memory that is used by the system to be dumped, the resulting dump is incomplete. If no limit is provided, all of the available physical memory is dumped. See “Preparing a multi-volume dump on ECKD DASD” on page 67 and Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711 for details.</td>
</tr>
<tr>
<td>-n</td>
<td>--noninteractive</td>
<td>n/a</td>
<td>Suppresses all confirmation prompts (for example, when preparing a DASD or tape dump device).</td>
</tr>
<tr>
<td>-p &lt;parmfile&gt;[,&lt;parm_addr&gt;]</td>
<td>--parmfile=&lt;parmfile&gt;[,&lt;parm_addr&gt;]</td>
<td>n/a</td>
<td>In a boot configuration, specifies the location of a kernel parameter file. In a SCSI dump configuration, specifies the location of a parameter file with SCSI system dumper parameters (see “Preparing a dump device on a SCSI disk” on page 68). You can specify multiple sources of kernel or SCSI system dumper parameters. For more information, see “How SCSI system dumper parameters from different sources are combined” on page 71 and “How kernel parameters from different sources are combined” on page 61. The optional &lt;parm_addr&gt; specifies the memory address where the combined kernel parameter list is to be loaded at IPL time. This specification is ignored for SCSI dump configuration. SCSI system dumper parameters are always loaded to the default address 0x1000.</td>
</tr>
<tr>
<td>Command line short option</td>
<td>Command line long option</td>
<td>Explanation</td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>--------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>-P &lt;parameters&gt;</td>
<td>--parameters=&lt;parameters&gt;</td>
<td>In a boot configuration, specifies kernel parameters.</td>
<td></td>
</tr>
<tr>
<td>parameters=&lt;parameters&gt;</td>
<td></td>
<td>In a SCSI dump configuration, specifies SCSI system dumper parameters (see “Preparing a dump device on a SCSI disk” on page 68).</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Individual parameters are single keywords or have the form key=value, without spaces. If you provide multiple parameters, separate them with a blank and enclose them within single quotation marks (') or double quotation marks (&quot;&quot;).</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>You can specify multiple sources of kernel or SCSI system dumper parameters. For more information, see “How SCSI system dumper parameters from different sources are combined” on page 71 and “How kernel parameters from different sources are combined” on page 61.</td>
<td></td>
</tr>
<tr>
<td>-r &lt;ramdisk&gt;,&lt;initrd_addr&gt;</td>
<td>--ramdisk=&lt;ramdisk&gt;,&lt;initrd_addr&gt;</td>
<td>Specifies the location of the initial RAM disk (initrd) on the file system and, optionally, in memory after IPL. If you do not specify a memory address, <em>zipl</em> investigates the location of other components and calculates a suitable address for you.</td>
<td></td>
</tr>
<tr>
<td>ramdisk=&lt;ramdisk&gt;,&lt;initrd_addr&gt;</td>
<td></td>
<td>Specifies the target directory where <em>zipl</em> creates boot-relevant files. The boot loader is installed on the disk that contains the target directory. For a SCSI dump device, this partition must be formatted with a file system supported by the SCSI system dumper (for example, ext4).</td>
<td></td>
</tr>
<tr>
<td>-t &lt;directory&gt;</td>
<td>--target=&lt;directory&gt;</td>
<td>For logical boot devices, specifies the device node of the base device, either by using the standard device name or in form of the major and minor number, separated by a colon (:).</td>
<td></td>
</tr>
<tr>
<td>target=&lt;directory&gt;</td>
<td></td>
<td>See “Using base device parameters” on page 63 for details.</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>--targetbase=&lt;targetbase_node&gt;</td>
<td>For logical boot devices, specifies the bytes per block of the base device.</td>
<td></td>
</tr>
<tr>
<td>targetbase=&lt;targetbase_node&gt;</td>
<td></td>
<td>See “Using base device parameters” on page 63 for details.</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>--targetblocksize=&lt;targetblocksize&gt;</td>
<td>For logical boot devices that map to ECKD type base devices, specifies the disk geometry of the base device in cylinders, heads, and sectors.</td>
<td></td>
</tr>
<tr>
<td>targetblocksize=&lt;targetblocksize&gt;</td>
<td></td>
<td>See “Using base device parameters” on page 63 for details.</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>--targetgeometry=&lt;cylinders&gt;,&lt;heads&gt;,&lt;sectors&gt;</td>
<td>For logical boot devices, specifies the offset in blocks between the start of the physical device and the start of the logical device.</td>
<td></td>
</tr>
<tr>
<td>targetgeometry=&lt;cylinders&gt;,&lt;heads&gt;,&lt;sectors&gt;</td>
<td></td>
<td>See “Using base device parameters” on page 63 for details.</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>--targetoffset=&lt;targetoffset&gt;</td>
<td>For logical boot devices, specifies the device type of the base device.</td>
<td></td>
</tr>
<tr>
<td>targetoffset=&lt;targetoffset&gt;</td>
<td></td>
<td>See “Using base device parameters” on page 63 for details.</td>
<td></td>
</tr>
<tr>
<td>Command line short option</td>
<td>Command line long option</td>
<td>Explanation</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>--------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>-T</td>
<td>--tape=</td>
<td>Specifies the tape device where <code>zipl</code> installs the boot loader code.</td>
<td></td>
</tr>
<tr>
<td>--tape=</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tape=</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **-v**
  
  --version
  
  n/a

- **-V**
  
  --verbose
  
  n/a

If you call `zipl` in configuration file mode without specifying a configuration file, the default `/etc/zipl.conf` is used. You can change the default configuration file with the environment variable `ZIPLCONF`.

### Configuration file structure

A configuration file comprises a default section and one or more sections with IPL configurations. In addition, there can be sections that define menu configurations.

- **[defaultboot]**
  
  A default section that defines what is to be done if the configuration file is called without a section specification.

- **[<configuration>]**
  
  One or more sections that describe IPL configurations.

- **:<menu_name>**
  
  Optionally, one or more menu sections that describe menu configurations.

A configuration file section consists of a section identifier and one or more option lines. Option lines are valid only as part of a section. Blank lines are permitted, and lines that begin with the number sign (#) are treated as comments and ignored. Option specifications consist of keyword=value pairs. There can but need not be blanks before and after the equal sign (=) of an option specification.

### Default section

The default section consists of the section identifier, **[defaultboot]**, followed by a single option line.

The option line specifies one of these mutually exclusive options:

- `default=<section_name>`
  
  Where `<section_name>` is one of the IPL configurations described in the configuration file. If the configuration file is called without a section specification, an IPL device is prepared according to this IPL configuration.
defaultmenu=<menu_name>

where <menu_name> is the name of a menu configuration that is described in the configuration file. If the configuration file is called without a section specification, IPL devices are prepared according to this menu configuration.

Examples

- This default specification points to a boot configuration boot1 as the default.
  ```
  [defaultboot]
  default=boot1
  ```
- This default specification points to a menu configuration with a menu menu1 as the default.
  ```
  [defaultboot]
  defaultmenu=menu1
  ```

IPL configurations

An IPL configuration has a section identifier that consists of a section name within square brackets and is followed by one or more option lines.

Each configuration includes one of the following mutually exclusive options that determine the type of IPL configuration:

- **image=<image>**
  Defines a boot configuration. See “Preparing a boot device” on page 59 for details.

- **dumpto=<dump_device>**
  Defines a DASD or tape dump configuration. See “Preparing a DASD or tape dump device” on page 65 for details.

- **mvdump=<dump_device_list>**
  Defines a multi-volume DASD dump configuration. See “Preparing a multi-volume dump on ECKD DASD” on page 67 for details.

- **dumptofs=<dump_partition>**
  Defines a SCSI dump configuration. See “Preparing a dump device on a SCSI disk” on page 68 for details.

Additional parameters might be required for logical boot devices (see “Preparing a logical device as a boot device” on page 62).

Menu configurations

For DASD and SCSI devices, you can define a menu configuration. A menu configuration has a section identifier that consists of a menu name with a leading colon.

The identifier is followed by one or more lines with references to IPL configurations in the same configuration file and one or more option lines.

- **target=<directory>**
  specifies a device where a boot loader is installed that handles multiple IPL configurations. For menu configurations, the target options of the referenced IPL configurations are ignored.

- **<i>=<configuration>**
  specifies a menu item. A menu includes one and more lines that specify the menu items.
<configuration> is the name of an IPL configuration that is described in the same configuration file. You can specify multiple boot configurations. For SCSI target devices, you can also specify one or more SCSI dump configurations. You cannot include DASD dump configurations as menu items.

<i> is the configuration number. The configuration number sequentially numbers the menu items, beginning with 1 for the first item. When initiating an IPL from a menu configuration, you can specify the configuration number of the menu item you want to use.

**default=<n>**

specifies the configuration number of one of the configurations in the menu to define it as the default configuration. If this option is omitted, the first configuration in the menu is the default configuration.

**prompt=<flag>**

for a DASD target device, determines whether the menu is displayed when an IPL is performed. Menus cannot be displayed for SCSI target devices.

For prompt=1 the menu is displayed, for prompt=0 it is suppressed. If this option is omitted, the menu is not displayed. Independent of this parameter, the operator can force a menu to be displayed by specifying prompt in place of a configuration number for an IPL configuration to be used.

If the menu of a menu configuration is not displayed, the operator can either specify the configuration number of an IPL configuration or the default configuration is used.

**timeout=<seconds>**

for a DASD target device and a displayed menu, specifies the time in seconds, after which the default configuration is IPLed, if no configuration has been specified by the operator. If this option is omitted or if 0 is specified as the timeout, the menu stays displayed indefinitely on the operator console and no IPL is performed until the operator specifies an IPL configuration.

**Example**

[Figure 21 on page 79](#) shows a sample configuration file that defines multiple configuration sections and two menu configurations.
The following commands assume that the configuration file of the sample is the default configuration file.
• Call `zipl` to use the default configuration file settings:

```
# zipl
```

**Result:** `zipl` reads the default option from the `[defaultboot]` section and selects the :menu1 section. It then installs a menu configuration with a boot configuration and a SCSI dump configuration.

• Call `zipl` to install a menu configuration (see also “Installing a menu configuration” on page 71):

```
# zipl -m menu2
```

**Result:** `zipl` selects the :menu2 section. It then installs a menu configuration with two DASD boot configurations. “Example for a DASD menu configuration on z/VM” on page 86 and “Example for a DASD menu configuration (LPAR)” on page 91 illustrate what this menu looks like when it is displayed.

• Call `zipl` to install a boot loader for boot configuration [boot2]:

```
# zipl boot2
```

**Result:** `zipl` selects the [boot2] section. It then installs a boot loader that loads copies of /boot/mnt/image-2, /boot/mnt/initrd, and /boot/mnt/parmf-2.

• Call `zipl` to prepare a tape that can be IPLed for a tape dump:

```
# zipl dumptape
```

**Result:** `zipl` selects the [dumptape] section and prepares a dump tape on /dev/rtibm0.

• Call `zipl` to prepare a DASD dump device:

```
# zipl dumpdasd -n
```

**Result:** `zipl` selects the [dumpdasd] section and prepares the dump device /dev/dasdc1. Confirmation prompts that require an operator response are suppressed.

• Call `zipl` to prepare a SCSI dump device:

```
# mount /dev/sda1 /boot
# mount /dev/sda2 /dumps
# mkdir /dumps/mydumps
# zipl dumpscsi
# umount /dev/sda1
# umount /dev/sda2
```

**Result:** `zipl` selects the [dumpscsi] section and prepares the dump device /dev/sda1. The associated dump file is created uncompressed in directory /mydumps on the dump partition. If space is required, the lowest-numbered dump file in the directory is deleted.
Chapter 6. Booting Linux

The options and requirements you have for booting Linux depend on your platform, LPAR or z/VM, and on your boot medium.

Find a general overview of how to boot Linux in an LPAR or in a z/VM guest virtual machine. For details about defining a Linux virtual machine, see z/VM Getting Started with Linux on System z®, SC24-6194, the chapter about creating your first Linux virtual machine.

IPL and booting

On IBM Z, you usually start booting Linux by performing an Initial Program Load (IPL).

Figure 22 summarizes the main steps of the boot process.

The IPL process accesses the IPL device and loads the Linux boot loader code to the mainframe memory. The boot loader code then gets control and loads the Linux kernel. At the end of the boot process Linux gets control.

If your Linux instance is to run in an LPAR, you can circumvent the IPL and use the service element (SE) to copy the Linux kernel to the mainframe memory (see "Loading Linux from removable media or from an FTP server" on page 95).

Apart from starting a boot process, an IPL can also be used for writing out system storage (dumping).
See Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711 for more information about dumps.

You can find the newest version of this book at [www.ibm.com/support/knowledgecenter/linuxonibm/lnz_r_distlibs.html](http://www.ibm.com/support/knowledgecenter/linuxonibm/lnz_r_distlibs.html)

Use the `zipl` tool to prepare DASD, SCSI, and tape devices as IPL devices for booting Linux or for dumping. For more information about `zipl`, see Chapter 5, “Initial program loader for IBM Z - zipl,” on page 55.

### Control point and boot medium

The control point from where you can start the boot process depends on the environment where your Linux is to run.

If your Linux instance is to run in LPAR mode, the control point is the mainframe's Support Element (SE) or an attached Hardware Management Console (HMC). For Linux on z/VM, the control point is the control program (CP) of the hosting z/VM.

The media that can be used as boot devices also depend on where Linux is to run. Table 11 provides an overview of the possibilities:

<table>
<thead>
<tr>
<th></th>
<th>DASD</th>
<th>tape</th>
<th>SCSI</th>
<th>z/VM reader</th>
<th>CD-ROM/FTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>z/VM guest</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>LPAR</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
</tbody>
</table>

DASDs, tapes on channel-attached tape devices, and SCSI device that are attached through an FCP channel can be used for both LPAR and z/VM guest virtual machines. A SCSI device can be a disk or an FC-attached CD-ROM or DVD drive. The z/VM reader is available only in a z/VM environment.

If your Linux runs in LPAR mode, you can also boot from a CD-ROM drive on the SE or HMC, or you can obtain the boot data from a remote FTP server.

### Menu configurations

In Red Hat Enterprise Linux 7.5, you can use `zipl` to prepare a DASD or SCSI boot disk with a menu configuration.

A boot device with a menu configuration can hold the code for multiple boot configurations. For SCSI disks, the menu can also include one or more SCSI system dumpers.

Each boot and dump configuration in a menu is associated with a configuration number. At IPL time, you can specify a configuration number to select the configuration to be used.

For menu configurations on DASD, you can display a menu with the configuration numbers (see “Example for a DASD menu configuration on z/VM” on page 86 and “Example for a DASD menu configuration (LPAR)” on page 91). For menu configurations on SCSI disks, you need to know the configuration numbers without being able to display the menus.
To boot Linux, you generally need a kernel image, boot loader code, kernel parameters, and an initial RAM disk image.

For sequential I/O boot devices, z/VM reader and tape, the order in which this data is provided is significant. For random access devices, there is no required order.

On Red Hat Enterprise Linux 7.5, kernel images are installed into the /boot directory and are named vmlinuz-<version>.s390x. For information about where to find the images and how to start an installation, see Red Hat Enterprise Linux 7.5 Installation Guide.

Boot loader code

Red Hat Enterprise Linux 7.5 kernel images are compiled to contain boot loader code for IPL from z/VM reader devices.

If you want to boot a kernel image from a device that does not correspond to the included boot loader code, you can provide alternate boot loader code separate from the kernel image.

Use zipl to prepare boot devices with separate DASD, SCSI, or tape boot loader code. You can then boot from DASD, SCSI, or tape regardless of the boot loader code in the kernel image.

Kernel parameters

The kernel parameters are in form of an ASCII text string of up to 895 characters. If the boot device is tape or the z/VM reader, the string can also be encoded in EBCDIC.

Individual kernel parameters are single keywords or keyword/value pairs of the form keyword=<value> with no blank. Blanks are used to separate consecutive parameters.

If you use the zipl command to prepare your boot device, you can provide kernel parameters on the command line, in a parameter file, and in a zipl configuration file.


Initial RAM disk image

An initial RAM disk holds files, programs, or modules that are not included in the kernel image but are required for booting.

Red Hat Enterprise Linux 7.5 provides a ramdisk in /boot and named initramfs-<kernel version>.s390x.img. When a ramdisk is installed or modified, you must call zipl to update the boot record.
Rebuilding the initial RAM disk image

Configuration changes might apply to components that are required in the boot process before the root file system is mounted.

For Red Hat Enterprise Linux 7, such components and their configuration are provided through an initial RAM disk.

Procedure

Perform these steps to make configuration changes for components in the initrd take effect:
1. Issue `dracut -f` to update the initial RAM disk of your target kernel.
2. Issue `zipl` to rewrite the zipl boot record.

Booting Linux in a z/VM guest virtual machine

You boot Linux in a z/VM guest virtual machine by issuing CP commands from a CMS or CP session.

For more general information about z/VM guest environments for Linux, see z/VM Getting Started with Linux on System z, SC24-6194.

Booting from a tape device

Boot Linux by issuing the IPL command with a tape boot device. The boot data on the tape must be arranged in a specific order.

Before you begin

You need a tape that is prepared as a boot device. A tape boot device must contain the following items in the specified order:
1. Tape boot loader code
   The tape boot loader code is included in the s390utils RPM.
2. Tape mark
3. Kernel image
4. Tape mark
5. Kernel parameters (optional)
6. Tape mark
7. Initial RAM disk (optional)
8. Tape mark
9. Tape mark

All tape marks are required even if an optional item is omitted. For example, if you do not provide an initial RAM disk image, the end of the boot information is marked with three consecutive tape marks. `zipl` prepared tapes conform to this layout. See “Preparing a boot device” on page 59 for information about preparing a tape with zipl.

Procedure

Perform these steps to start the boot process:
1. Establish a CMS or CP session with the z/VM guest virtual machine where you want to boot Linux.
2. Ensure that the boot device is accessible to your z/VM guest virtual machine.
3. Ensure that the correct tape is inserted and rewound.

4. Issue a command of this form:

   ```
   #cp i <devno> clear parm <kernel_parameters>
   ```

   where

   `<devno>`
   is the device number of the boot device as seen by the guest virtual machine.

   `parm <kernel_parameters>`
   is an optional 64-byte string of kernel parameters to be concatenated to the end of the existing kernel parameters that are used by your boot configuration (see “Preparing a boot device” on page 59 for information about the boot configuration).

   See also “Specifying kernel parameters when booting Linux” on page 25.

**Booting from a DASD**

**Before you begin**

You need a DASD boot device that is prepared with zipl (see “Preparing a boot device” on page 59).

**Procedure**

Perform these steps to start the boot process:

1. Establish a CMS or CP session with the z/VM guest virtual machine where you want to boot Linux.

2. Ensure that the boot device is accessible to your z/VM guest virtual machine.

3. Issue a command of this form:

   ```
   #cp i <devno> clear loadparm <n> parm <kernel_parameters>
   ```

   where:

   `<devno>`
   specifies the device number of the boot device as seen by the guest.

   `loadparm <n>`
   is applicable to menu configurations only. Omit this parameter if you are not working with a menu configuration.

   Configuration number 0 specifies the default configuration. Depending on the menu configuration, omitting this option might display the menu or select the default configuration. Specifying “prompt” instead of a configuration number forces the menu to be displayed.

   When the menu is displayed, you can specify additional kernel parameters (see “Example for a DASD menu configuration on z/VM” on page 86). These additional kernel parameters are appended to the parameters you might have provided in a parameter file. The combined parameter string must not exceed 895 bytes.

   See “Menu configurations” on page 77 for more details about menu configurations.
parm <kernel_parameters>

is an optional 64-byte string of kernel parameters to be concatenated to the end of the existing kernel parameters that are used by your boot configuration (see “Preparing a boot device” on page 59 for information about the boot configuration).

See also “Specifying kernel parameters when booting Linux” on page 25.

Example for a DASD menu configuration on z/VM

Use the VI VMSG z/VM CP command to choose a boot configuration from a menu configuration.

This example illustrates how menu2 in the sample configuration file in Figure 21 on page 79 is displayed on the z/VM guest virtual machine console:

```
00: zipl interactive boot menu
00: 0. default (boot1)
00: 1. boot1
00: 2. boot3
00: Note: VM users please use ’#cp vi vmsg <number> <kernel-parameters>’
00:
00: Please choose (default will boot in 30 seconds): #cp vi vmsg 2
```

You choose a configuration by specifying the configuration number. For example, to boot configuration boot3 specify

```
#cp vi vmsg 2
```

You can also specify additional kernel parameters by appending them to the configuration number. For example, you can specify:

```
#cp vi vmsg 2 maxcpus=1
```

These parameters are concatenated to the end of the existing kernel parameters that are used by your boot configuration when booting Linux.

Booting from a SCSI device

Boot Linux by issuing the IPL command with an FCP channel as the IPL device. You must specify the target port and LUN for the boot device in advance by setting the z/VM CP LOADDEV parameter.

Before you begin

You need a SCSI boot device that is prepared with zipl (see “Preparing a boot device” on page 59). A SCSI device can be a disk or an FC-attached CD-ROM or DVD drive.

Procedure

Perform these steps to start the boot process:

1. Establish a CMS or CP session with the z/VM guest virtual machine where you want to boot Linux.
2. Ensure that the FCP channel that provides access to the SCSI boot disk is accessible to your z/VM guest virtual machine.
3. Specify the target port and LUN of the SCSI boot disk. Enter a command of this form:
where:

<wwpn>

specifies the world wide port name (WWPN) of the target port in hexadecimal format. A blank separates the first eight digits from the final eight digits.

<lun>

specifies the LUN of the SCSI boot disk in hexadecimal format. A blank separating the first eight digits from the final eight digits.

**Example:** To specify a WWPN 0x5005076300c20b8e and a LUN 0x5241000000000000:

```
#cp set loaddev portname 50050763 00c20b8e lun 52410000 00000000
```

4. Optional for menu configurations: Specify the boot configuration (boot program in z/VM terminology) to be used. Enter a command of this form:

```
#cp set loaddev bootprog <n>
```

where <n> specifies the configuration number of the boot configuration.

Omitting the bootprog parameter or specifying the value 0 selects the default configuration. For more information about menu configurations, see "Menu configurations" on page 77.

**Example:** To select a configuration with configuration number 2 from a menu configuration:

```
#cp set loaddev bootprog 2
```

5. Optional: Specify kernel parameters.

```
#cp set loaddev scpdata <APPEND|NEW> '<kernel_parameters>'
```

where:

<kernel_parameters>

specifies a set of kernel parameters to be stored as system control program data (SCPDATA). When booting Linux, these kernel parameters are concatenated to the end of the existing kernel parameters that are used by your boot configuration.

<kernel_parameters> must contain ASCII characters only. If characters other than ASCII characters are present, the boot process ignores the SCPDATA.

<kernel_parameters> as entered from a CMS or CP session is interpreted as lowercase on Linux. If you require uppercase characters in the kernel parameters, run the SET LOADDEV command from a REXX script instead. In the REXX script, use the “address command” statement. See REXX/VM Reference, SC24-6221 and REXX/VM User’s Guide, SC24-6222 for details.

**Optional:** APPEND

appends kernel parameters to existing SCPDATA. This is the default.
Optional: NEW
replaces existing SCPDATA.

Examples:

- To append kernel parameter noresume to the current SCPDATA:
  ```
  # cp set loaddev scpdata 'noresume'
  ```

- To replace the current SCPDATA with the kernel parameters
  `resume=/dev/sda2 and no_console_suspend`:
  ```
  # cp set loaddev scpdata NEW 'resume=/dev/sda2 no_console_suspend'
  ```

For a subsequent IPL command, these kernel parameters are concatenated to
the end of the existing kernel parameters in your boot configuration.

6. Start the IPL and boot process by entering a command of this form:
   ```
   # cp i <devno>
   ```
   
   where `<devno>` is the device number of the FCP channel that provides access
to the SCSI boot disk.

Tip

You can specify the target port and LUN of the SCSI boot disk, a boot
configuration, and SCPDATA all with a single SET LOADDEV command. See
`z/VM CP Commands and Utilities Reference, SC24-6175` for more information about
the SET LOADDEV command.

Booting from the z/VM reader

Boot Linux by issuing the IPL command with the z/VM reader as the IPL device.
You first must transfer the boot data to the reader.

Before you begin

You need the following files, all in record format fixed 80:

- Linux kernel image with built-in z/VM reader boot loader code. This is the case
  for the default Red Hat Enterprise Linux 7.5 kernel.
- Kernel parameters (optional)
- Initial RAM disk image (optional)

About this task

This information is a summary of how to boot Linux from a z/VM reader. For
more details, refer to Redpaper™ `Building Linux Systems under IBM VM, REDP-0120`.

Procedure

Proceed like this to boot Linux from a z/VM reader:

1. Establish a CMS session with the guest where you want to boot Linux.
2. Transfer the kernel image, kernel parameters, and the initial RAM disk image
to your guest. You can obtain the files from a shared minidisk or use:
   - The z/VM sendfile facility.
• An FTP file transfer in binary mode.

Files that are sent to your reader contain a file header that you must remove before you can use them for booting. Receive files that you obtain through your z/VM reader to a minidisk.

3. Set up the reader as a boot device.
   a. Ensure that your reader is empty.
   b. Direct the output of the punch device to the reader. Issue:

   \[\text{spool pun * rdr}\]

   c. Use the CMS PUNCH command to transfer each of the required files to the reader. Be sure to use the “no header” option to omit the file headers.

   First transfer the kernel image.
   Second transfer the kernel parameters.
   Third transfer the initial RAM disk image, if present.

   For each file, issue a command of this form:

   \[\text{pun <file_name> <file_type> <file_mode>} (noh}\]

   d. Optional: Ensure that the contents of the reader remain fixed.

   \[\text{change rdr all keep nohold}\]

   If you omit this step, all files are deleted from the reader during the IPL that follows.

4. Issue the IPL command:

   \[\text{ipl 000c clear parm <kernel_parameters>}\]

   where:

   \[0x000c\]
   is the device number of the reader.

   \[\text{parm <kernel_parameters>}\]
   is an optional 64-byte string of kernel parameters to be concatenated to the end of the existing kernel parameters that are used by your boot configuration (see \[\text{Preparing a boot device" on page 59}\] for information about the boot configuration).

   See also \["Specifying kernel parameters when booting Linux” on page 25\].

---

**Booting Linux in LPAR mode**

You can boot Linux in LPAR mode from a Hardware Management Console (HMC) or Support Element (SE).

**About this task**

The following description refers to an HMC, but the same steps also apply to an SE.

**Booting from DASD**

Use the SE or HMC to boot Linux in LPAR mode from a DASD boot device.
Before you begin

You need a boot device that is prepared with `zip` (see "Preparing a boot device" on page 59).

Procedure

Perform these steps to boot from a DASD boot device:

1. In the navigation pane of the HMC, expand **Systems Management** and **Servers** and select the mainframe system you want to work with. A table of LPARs is displayed on the **Images** tab in the content area.
2. Select the LPAR where you want to boot Linux.
3. In the **Tasks** area, expand **Recovery** and click **Load** (see Figure 23).
4. Select the load type **Normal** (see Figure 24 on page 91).
5. Enter the device number of the DASD boot device in the **Load address** field.

6. If the boot configuration is part of a **zipl** created menu configuration, enter the configuration number that identifies your DASD boot configuration within the menu in the **Load parameter** field.

   Configuration number **0** specifies the default configuration. Depending on the menu configuration, omitting this option might display the menu or select the default configuration. Specifying “prompt” instead of a configuration number forces the menu to be displayed.

   When the menu is displayed, you can specify additional kernel parameters (see “Example for a DASD menu configuration (LPAR)” in the configuration file). These additional kernel parameters are appended to the parameters you might have provided in a parameter file. The combined parameter string must not exceed 895 bytes. See “Menu configurations” on page 77 for more details about menu configurations.

7. Click **OK** to start the boot process.

**Results**

Check the output on the preferred console (see “Console kernel parameter syntax” on page 40) to monitor the boot progress.

**Example for a DASD menu configuration (LPAR)**

This example illustrates how menu2 in the sample configuration file in Figure 21 on page 79 is displayed on the HMC or SE:
You choose a configuration by specifying the configuration number. For example, to boot configuration boot3, issue:

```
# 2
```

You can also specify additional kernel parameters by appending them to this command. For example:

```
# 2 maxcpus=1
```

### Booting from SCSI

Use the SE or HMC to boot Linux in LPAR from a SCSI boot device.

#### Before you begin

You need a boot device that is prepared with `zipl` (see "Preparing a boot device" on page 59).

#### Procedure

Perform these steps to boot from a SCSI boot device:

1. In the navigation pane of the HMC, expand **Systems Management** and **Servers** and select the mainframe system you want to work with. A table of LPARs is displayed on the **Images** tab in the content area.
2. Select the LPAR where you want to boot Linux.
3. In the **Tasks** area, expand **Recovery** and click **Load** (see Figure 25 on page 93).
4. A SCSI device can be a disk or an FC-attached CD-ROM or DVD drive. Select load type SCSI (see Figure 26).

5. Enter the device number of the FCP channel through which the SCSI device is accessed in the Load address field.

6. Enter the WWPN of the SCSI device in the World wide port name field.
7. Enter the LUN of the SCSI device in the **Logical unit number** field.

8. If the boot configuration is part of a `zipl` created menu configuration, enter the configuration number that identifies your SCSI boot configuration within the menu in the **Boot program selector** field. Configuration number 0 specifies the default configuration. For example, an installation from DVD is typically done with boot program selector 2.

   See “Menu configurations” on page 77 for more details about menu configurations.

9. Optional: Type kernel parameters in the **Operating system specific load parameters** field. These parameters are concatenated to the end of the existing kernel parameters that are used by your boot configuration when booting Linux.

   Use ASCII characters only. If you enter characters other than ASCII characters, the boot process ignores the data in the **Operating system specific load parameters** field.

10. Accept the defaults for the remaining fields.

11. Click **OK** to start the boot process.

**Results**

Check the output on the preferred console (see “Console kernel parameter syntax” on page 40) to monitor the boot progress.

**Booting from tape**

You can boot Linux in LPAR mode from tape.

**Before you begin**

You need a boot device that is prepared with `zipl` (see “Preparing a boot device” on page 59).

**Procedure**

Perform these steps to boot from a tape boot device:

1. In the navigation pane of the HMC, expand **Systems Management** and **Servers** and select the mainframe system you want to work with. A table of LPARs is displayed on the **Images** tab in the content area.

2. Select the LPAR where you want to boot Linux.

3. In the **Tasks** area, expand **Recovery** and click **Load** (see Figure 27 on page 95).
4. Select load type Normal (see Figure 24 on page 91).
5. Enter the device number of the tape boot device in the **Load address** field.
6. Click **OK** to start the boot process.

**Results**

Check the output on the preferred console (see “Console kernel parameter syntax” on page 40) to monitor the boot progress.

**Loading Linux from removable media or from an FTP server**

Instead of a boot loader, you can use SE functions to copy the Linux kernel image to your LPAR memory.

After the Linux kernel is loaded, Linux is started using restart PSW.

**Before you begin**

You need installation data that includes a special file with installation information (with extension “.ins”). This file can be in different locations:

- On a disk that is inserted in the CD-ROM or DVD drive of the system where the HMC runs
- In the file system of an FTP server that you can access through FTP from your HMC system

The .ins file contains a mapping of the location of installation data on the disk or FTP server and the memory locations where the data is to be copied.

For Red Hat Enterprise Linux 7.5, this file is called generic.ins and in the root directory of the file system on the DVD.
Procedure

Perform these steps:

1. In the navigation pane of the HMC expand **Systems Management** and **Servers** and select the mainframe system that you want to work with. A table of LPARs is displayed on the **Images** tab in the content area.

2. Select the LPAR where you want to boot Linux.

3. In the **Tasks** area, expand **Recovery** and click **Load from Removable Media or Server** (see Figure 28).

4. Specify the source of the code to be loaded.
   - For loading from a CD-ROM drive:
     a. Select **Hardware Management Console CD-ROM/DVD** (see Figure 29 on page 97).

![Figure 28. Load from Removable Media or Server task on the HMC](image-url)
b. Leave the **File location** field blank.

- For loading from an FTP server:
  a. Select **FTP Source**.
  b. Enter the IP address or host name of the FTP server with the installation code in the **Host computer** entry field.
  c. Enter your user ID for the FTP server in the **User ID** entry field.
  d. Enter your password for the FTP server in the **Password** entry field.
  e. If required by your FTP server, enter your account information in the **Account** entry field.
  f. Enter the path to the directory with the `generic.ins` in the file location entry field. You can leave this field blank if the file is in the FTP server's root directory.

5. Click **Continue** to display the Select Software to Install panel (Figure 30).


7. Click **OK** to start loading Linux.
Results

The kernel has started and the Red Hat Enterprise Linux 7.5 boot process continues.

Displaying current IPL parameters

To display the IPL parameters, use the `lsreipl` command with the `-i` option. Alternatively, a sysfs interface is available.

For more information about the `lsreipl` command, see "lsreipl - List IPL and re-IPL settings" on page 583. In sysfs, information about IPL parameters is available in subdirectories of `/sys/firmware/ipl`:

```
/sys/firmware/ipl/ipl_type
```

The `/sys/firmware/ipl/ipl_type` file contains the device type from which the kernel was booted. The following values are possible:

- **ccw** The IPL device is a CCW device, for example, a DASD or the z/VM reader.
- **fcp** The IPL device is an FCP device.
- **unknown** The IPL device is not known.

Depending on the IPL type, there might be more files in `/sys/firmware/ipl/`.

If the device is a CCW device, the additional files `device` and `loadparm` are present.

```
device
```

Contains the bus ID of the CCW device that is used for IPL, for example:

```
# cat /sys/firmware/ipl/device
0.0.1234
```

```
loadparm
```

Contains up to 8 characters for the loadparm that is used for IPL, for example:

```
# cat /sys/firmware/ipl/loadparm
1
```

```
parm
```

Contains additional kernel parameters that are specified with the PARM parameter when booting with the z/VM CP IPL command, for example:

```
# cat /sys/firmware/ipl/parm
noresume
```

See also "Specifying kernel parameters when booting Linux" on page 25.

A leading equal sign (=) indicates that the existing kernel parameters used by the boot configuration were ignored and the kernel parameters of the `parm` attribute where the only kernel parameters used for booting Linux.

See "Replacing all kernel parameters in a boot configuration" on page 26.

If the device is FCP, a number of additional files are present (also see Chapter 11, "SCSI-over-Fibre Channel device driver," on page 159 for details):
**device**  
Contains the bus ID of the FCP device that is used for IPL, for example:

```
# cat /sys/firmware/ipl/device
0.0.50dc
```

**wwpn**  
Contains the WWPN used for IPL, for example:

```
# cat /sys/firmware/ipl/wwpn
0x5005076300c20b8e
```

**lun**  
Contains the LUN used for IPL, for example:

```
# cat /sys/firmware/ipl/lun
0x5010000000000000
```

**br_lba**  
Contains the logical block address of the boot record on the boot device (usually 0).

**bootprog**  
Contains the boot program number.

**scp_data**  
Contains additional kernel parameters that are used when booting from a SCSI device (see “Boot from a SCSI device” on page 86 and “Boot from DASD” on page 89). A leading equal sign (=) indicates that the existing kernel parameters used by the boot configuration were ignored and the kernel parameters of the scp_data attribute were the only kernel parameters used for booting Linux.

```
# cat /sys/firmware/ipl/scp_data
noresume
```

**binary_parameter**  
Contains the information of the preceding files in binary format.

---

### Rebooting from an alternative source

When you reboot Linux, the system conventionally boots from the last used location. However, you can configure an alternative device to be used for re-IPL instead of the last used IPL device.

When the system is re-IPLed, the alternative device is used to boot the kernel.

To configure the re-IPL device, use the `chreipl` tool (see “chreipl - Modify the re-IPL configuration” on page 497).

Alternatively, you can use a sysfs interface. The virtual configuration files are located under `/sys/firmware/reipl`. To configure, write strings into the configuration files. The following re-IPL types can be set with the `/sys/firmware/reipl/reipl_type` attribute:

- **ccw**  
  For ccw devices such as DASDs that are attached through ESCON or FICON®.

- **fcp**  
  For FCP SCSI devices, including SCSI disks and CD or DVD drives (Hardware support is required.)

- **nss**  
  For Named Saved Systems (z/VM only)
For each supported re-IPL type a sysfs directory is created under 
/sys/firmware/reipl that contains the configuration attributes for the device. The
directory name is the same as the name of the re-IPL type.

When Linux is booted, the re-IPL attributes are set by default to the values of the
boot device, which can be found under /sys/firmware/ipl.

**Attributes for ccw**

You can find the attributes for re-IPL type ccw in the /sys/firmware/reipl/ccw
sysfs directory.

**device**  
Device number of the re-IPL device. For example 0.0.7412 or 0.1.5119.

**loadparm**  
Up to eight characters for the loadparm used to select the boot
configuration in the zipl menu (if available).

**parm**  
A 64-byte string of kernel parameters that is concatenated to the boot
command-line. The PARM parameter can be set only for Linux on z/VM. See also “Specifying kernel parameters when booting Linux” on page 25.

A leading equal sign (=) means that the existing kernel parameter line in
the boot configuration is ignored and the boot process uses the kernel
parameters in the parm attribute only. See also “Replacing all kernel
parameters in a boot configuration” on page 26.

**Attributes for fcp**

You can find the attributes for re-IPL type fcp in the /sys/firmware/reipl/fcp
sysfs directory.

**device**  
Device number of the FCP device that is used for re-IPL. For example, 0.0.7412.

**wwpn**  
World wide port number of the FCP re-IPL device.

**lun**  
Logical unit number of the FCP re-IPL device.

**bootprog**  
Boot program selector. Used to select the boot configuration in the zipl
menu (if available).

**br_lba**  
Boot record logical block address. Master boot record. Is always 0 for
Linux.

**scp_data**  
Kernel parameters to be used for the next FCP re-IPL.

A leading equal sign (=) means that the existing kernel parameter line in
the boot configuration is ignored and the boot process uses the kernel
parameters in the scp_data attribute only. See also “Replacing all kernel
parameters in a boot configuration” on page 26.

**Attributes for nss**

You can find the attributes for re-IPL type nss in the /sys/firmware/reipl/nss
sysfs directory.

**name**  
Name of the NSS. The NSS name can be 1-8 characters long and must
consist of alphabetic or numeric characters. The following examples are all valid NSS names: 73248734, NSSCSITE, or NSS1234.

**parm** A 56-byte string of parameters that is passed to the NSS to be booted.

### Kernel panic settings

Set the attribute `/sys/firmware/shutdown_actions/on_panic` to `reipl` to make the system re-IPL with the current re-IPL settings if a kernel panic occurs.


### Examples for configuring re-IPL

Typical examples include configuring re-IPL from an FCP device and specifying parameters for re-IPL:

- To configure an FCP re-IPL device 0.0.5711 with a LUN 0x1711000000000000 and a WWPN 0x5005076303004715 with an additional kernel parameter `noresume`:

```
# echo 0.0.5711 > /sys/firmware/reipl/fcp/device
# echo 0x5005076303004715 > /sys/firmware/reipl/fcp/wwpn
# echo 0x1711000000000000 > /sys/firmware/reipl/fcp/lun
# echo 0 > /sys/firmware/reipl/fcp/bootprog
# echo 0 > /sys/firmware/reipl/fcp/br_lba
# echo "noresume" > /sys/firmware/reipl/fcp/scp_data
# echo fcp > /sys/firmware/reipl/reipl_type
```

- To specify additional kernel parameters for Linux re-IPL, follow these steps:
  1. Change to the sysfs directory appropriate for the next re-IPL:

```
# cd /sys/firmware/reipl/$(cat /sys/firmware/reipl/reipl_type)/sys/firmware/reipl/ccw
```

  2. Use the `echo` command to output the parameter string into the `parm` attribute:

```
# echo "noresume" > parm
```
Chapter 7. Suspending and resuming Linux

With suspend and resume support, you can stop a running instance of Red Hat Enterprise Linux 7.5 and later continue operations.

When Linux is suspended, data is written to a swap partition. The resume process uses this data to make Linux continue from where it left off when it was suspended. A suspended Linux instance does not require memory or processor cycles.

Suspend and resume support applies to both Linux on z/VM and Linux instances that run directly in an LPAR.

While a Linux instance is suspended, you can run another Linux instance in the z/VM guest virtual machine or in the LPAR where the suspended Linux instance was running.

What you should know about suspend and resume

Before suspending a Linux instance, you must be aware of the prerequisites and of activities that can cause resume to fail.

Prerequisites for suspending a Linux instance

Suspend and resume support checks for conditions that might prevent resuming a suspended Linux instance. You cannot suspend a Linux instance unless all prerequisites are fulfilled.

The following prerequisites must be fulfilled regardless of whether a Linux instance runs directly in an LPAR or as a z/VM guest:

- All tape device nodes must be closed and online tape drives must be unloaded.
- The Linux instance must not have used any hotplug memory since it was last booted.
- No program must be in a prolonged uninterruptible sleep state. Programs can assume this state while they are waiting for an outstanding I/O request to complete. Most I/O requests complete in a very short time and do not compromise suspend processing. An example of an I/O request that can take too long to complete is rewinding a tape.

For Linux on z/VM, the following additional prerequisites must be fulfilled:

- No discontiguous saved segment (DCSS) device must be accessed in exclusive-writable mode.
  You must remove all DCSSs of segment types EW, SW, and EN by writing the DCSS name to the sysfs remove attribute.
  You must remove all DCSSs of segment types SR and ER that are accessed in exclusive-writable mode or change their access mode to shared.
  For more information, see "Removing a DCSS device" on page 409 and "Setting the access mode" on page 406.
- All device nodes of the z/VM recording device driver must be closed.
- All device nodes of the z/VM unit record device driver must be closed.
- No watchdog timer must run and the watchdog device node must be closed.
Precautions while a Linux instance is suspended

There are conditions outside the control of the suspended Linux instance that can cause resume to fail.

- The CPU configuration must remain unchanged between suspend and resume.
- The data that is written to the swap partition when the Linux instance is suspended must not be compromised.

In particular, be sure that the swap partition is not used if another operating system instance runs in the LPAR or z/VM guest virtual machine while the initial Linux instance is suspended.

- If the Linux instance uses expanded storage (XPRAM), this expanded storage must remain unchanged until the Linux instance is resumed.

If the size or content of the expanded memory is changed before the Linux instance is resumed or if the expanded memory is unavailable when the Linux instance is resumed, resuming fails with a kernel panic.

- If an instance of Linux on z/VM uses one or more DCSSs these DCSSs must remain unchanged until the Linux instance is resumed.

If the size, location, or content of a DCSS is changed before the Linux instance is resumed, resuming fails with a kernel panic.

- For an instance of Linux on z/VM with a Linux kernel that is a named saved system (NSS), the NSS must remain unchanged until the Linux instance is resumed.

If the size, location, or content of the NSS is changed before the Linux instance is resumed, resuming fails.

- Take special care when replacing a DASD and, thus, making a different device available at a particular device bus-ID.

You might intentionally replace a device with a backup device. Changing the device also changes its UID-based device nodes. Expect problems if you run an application that depends on UID-based device nodes and you exchange one of the DASD the application uses. In particular, you cannot use multipath tools when the UID changes.

- The SCSI configuration must remain unchanged until the Linux instance is resumed.

- Generally, avoid changes to the real or virtual hardware configuration between suspending and resuming a Linux instance.

- Disks that hold swap partitions or the root file system must be present when resuming the Linux instance.

Handling of devices that are unavailable when resuming

Devices that were available when the Linux instance was suspended might be unavailable when resuming.

If such unavailable devices were offline when the Linux instance was suspended, they are de-registered and the device name can be assigned to other devices.

If unavailable devices where online when the Linux instance was suspended, handling depends on the respective device driver. DASD and FCP devices remain registered as disconnected devices. The device name and the device configuration are preserved. Devices that are controlled by other device drivers are de-registered.
Handling of devices that become available at a different subchannel

The mapping between subchannels and device bus-IDs can change if the real or virtual hardware is restarted between suspending and resuming Linux.

If the subchannel changes for a DASD or FCP device, the device configuration is changed to reflect the new subchannel. This change is accomplished without de-registration. Thus, device name and device configuration are preserved.

If the subchannel changes for any other device, the device is de-registered and registered again as a new device.

Setting up Linux for suspend and resume

Configure suspend and resume support through kernel parameters and set up a suitable swap partition for suspending and resuming a Linux instance.

Kernel parameters

You configure the suspend and resume support by adding parameters to the kernel parameter line.

<table>
<thead>
<tr>
<th>suspend and resume kernel parameter syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>→ → resume=&lt;device_node&gt; [no_console_suspend] [noresume] ← ←</td>
</tr>
</tbody>
</table>

where:

- **resume=<device_node>**  
  specifies the standard device node of the swap partition with the data that is required for resuming the Linux instance.

- **no_console_suspend**  
  prevents Linux consoles from being suspended early in the suspend process. Without this parameter, you cannot see the kernel messages that are issued by the suspend process.

- **noresume**  
  boots the kernel without resuming a previously suspended Linux instance. Add this parameter to circumvent the resume process, for example, if the data written by the previous suspend process is damaged.

Example

To use a partition /dev/disk/by-path/ccw-0.0.2f50-part1 as the swap partition and prevent Linux consoles from being suspended early in the suspend process specify:

```
resume=/dev/disk/by-path/ccw-0.0.2f50-part1 no_console_suspend
```

The example uses a "by-path" definition of the disk to ensure that the correct device is used, for example, after a reboot.
Setting up a swap partition

During the suspend process, Linux writes data to a swap partition. This data is required later to resume Linux.

Set up a swap partition that is at least the size of the available LPAR memory or the memory of the z/VM guest virtual machine.

Do not use this swap partition for any other operating system that might run in the LPAR or z/VM guest virtual machine while the Linux instance is suspended.

You cannot suspend a Linux instance while most of the memory and most of the swap space is in use. If there is not sufficient remaining swap space to hold the data for resuming the Linux instance, suspending the Linux instance fails.

To assure sufficient swap space you might have to configure two swap partitions, one partition for regular swapping and another for suspending the Linux instance. Configure the swap partition for suspending the Linux instance with a lower priority than the regular swap partition.

Use the pri= parameter to specify the swap partitions in /etc/fstab with different priorities. See the swapon man page for details.

The following example shows two swap partitions with different priorities:

```
# cat /etc/fstab
... 
/dev/dasdb1 swap swap pri=-1 0 0
/dev/dasdc1 swap swap pri=-2 0 0
```

In the example, the partition to be used for the resume data is /dev/dasdc1.

You can check your current swap configuration by reading /proc/swaps.

```
# cat /proc/swaps
Filename      Type    Size    Used   Priority
/dev/dasdb1   partition 7212136 71056   -1
/dev/dasdc1   partition 7212136 0     -2
```

Configuring for fast resume

The more devices are available to a Red Hat Enterprise Linux instance, the longer it takes to resume a suspended instance.

With a thousand or more available devices, the resume process can take longer than an IPL. If the duration of the resume process is critical for a Linux instance with many devices, include unused devices in the exclusion list (see "cio_ignore-List devices to be ignored" on page 652).

Suspending a Linux instance

Suspend a Linux instance by writing to the /sys/power/state sysfs attribute.
Before you begin

Attention: Suspend only Linux instances for which you specified the \texttt{resume=} kernel parameter. Without this parameter, you cannot resume the suspended Linux instance.

Procedure

Enter the following command to suspend a Linux instance:

\begin{verbatim}
# echo disk > /sys/power/state
\end{verbatim}

Results

On the Linux console you might see progress indications until the console itself is suspended. Most of these messages require log level 7 or higher to be printed. See “Using the magic sysrequest feature” on page 49 about setting the log level. You cannot see such progress messages if you suspend the Linux instance from an ssh session.

Resuming a suspended Linux instance

Boot Linux to resume a suspended Linux instance.

About this task

Use the same kernel, initial RAM disk, and kernel parameters that you used to first boot the suspended Linux instance.

You must reestablish any terminal session for HVC terminal devices and for terminals that are provided by the iucvtty program. You also must reestablish all ssh sessions that timed out while the Linux instance was suspended.

If resuming the Linux instance fails, boot Linux again with the \texttt{noresume} kernel parameter. The boot process then ignores the data that was written to the swap partition and starts Linux without resuming the suspended instance.
Chapter 8. Shutdown actions

Several triggers can cause Linux to shut down. For each shutdown trigger, you can configure a specific shutdown action to be taken as a response.

**Table 12. Shutdown triggers and default action overview**

<table>
<thead>
<tr>
<th>Trigger</th>
<th>Command or condition</th>
<th>Default shutdown action</th>
</tr>
</thead>
<tbody>
<tr>
<td>halt</td>
<td>Linux <code>shutdown -H</code> command</td>
<td>stop</td>
</tr>
<tr>
<td>poff</td>
<td>Linux <code>poweroff</code> or <code>shutdown -P</code> command</td>
<td>stop</td>
</tr>
<tr>
<td>reboot</td>
<td>Linux <code>reboot</code> or <code>shutdown -r</code> command</td>
<td>reipl</td>
</tr>
</tbody>
</table>
| restart | • `PSW restart` on the HMC for Linux in LPAR mode  
          • `z/VM CP system restart` command for Linux on `z/VM` | stop                    |
| panic   | Linux kernel panic | stop                    |

The available shutdown actions are summarized in **Table 13**.

**Table 13. Shutdown actions**

<table>
<thead>
<tr>
<th>Action</th>
<th>Explanation</th>
<th>See also</th>
</tr>
</thead>
<tbody>
<tr>
<td>stop</td>
<td>For panic or restart, enters a disabled wait state. For all other shutdown triggers, stops all CPUs.</td>
<td>n/a</td>
</tr>
<tr>
<td>ipl</td>
<td>Performs an IPL according to the specifications in <code>/sys/firmware/ipl</code>.</td>
<td>“Displaying current IPL parameters” on page 98</td>
</tr>
<tr>
<td>reipl</td>
<td>Performs an IPL according to the specifications in <code>/sys/firmware/reipl</code>.</td>
<td>“Rebooting from an alternative source” on page 99</td>
</tr>
<tr>
<td>dump</td>
<td>Creates a dump according to the specifications in <code>/sys/firmware/dump</code>.</td>
<td>Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711</td>
</tr>
<tr>
<td>dump_reipl</td>
<td>Performs the dump action followed by the reipl action.</td>
<td></td>
</tr>
<tr>
<td>vmcmd</td>
<td>For Linux on <code>z/VM</code>, issues one or more <code>z/VM CP commands according to the specifications in /sys/firmware/vmcmd</code>.</td>
<td>“Configuring z/VM CP commands as a shutdown action” on page 112</td>
</tr>
</tbody>
</table>

Use `lsshut` to find out which shutdown action is configured for each shutdown trigger, see “`lsshut - List the current system shutdown actions`” on page 586.
Use the applicable command to configure the shutdown action for a shutdown trigger:

- For **halt**, **power off**, and **reboot** use `chshut`, see "chshut - Control the system shutdown actions" on page 501.
- For **restart** and **panic** use `dumpconf`, see Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711

**kdump for restart and panic**

If kdump is set up for a Linux instance, kdump is started to create a dump, regardless of the shutdown actions that are specified for restart and panic. With kdump, these settings act as a backup that is used only if kdump fails.

**Note:** kdump is not a shutdown action that you can set as a sysfs attribute value for a shutdown trigger. See Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711 about how to set up kdump.

**Override for halt**

Red Hat Enterprise Linux maps **halt** to **poff**. The action that is specified for **halt** is ignored and the action that is specified for **poff** is triggered instead. You can change this by editing the file `/etc/sysconfig/shutdown` and replacing HALT="auto" with HALT="halt".

### The shutdown configuration in sysfs

The configured shutdown action for each shutdown trigger is stored in a sysfs attribute `/sys/firmware/shutdown_actions/on_<trigger>`.

```
Figure 31. sysfs branch with shutdown action settings
```

The preferred way to read or change these settings is using the `1sshut`, `chshut`, and `dumpconf` commands. Alternatively, you can read and write to the `/sys/firmware/shutdown_actions/on_<trigger>` attributes.

**Examples**

- This command reads the shutdown setting for the **poff** shutdown trigger.
  ```
  # cat /sys/firmware/shutdown_actions/on_poff
  stop
  ```

- This command changes the setting for the **restart** shutdown trigger to **ipl**:
  ```
  # echo ipl > /sys/firmware/shutdown_actions/on_restart
  ```
Details for the ipl, reipl, dump, and vmcmd shutdown actions are contained in the corresponding subdirectories in /sys/firmware. For example, /sys/firmware/ipl contains specifications for an IPL device and other IPL parameters.
Configuring z/VM CP commands as a shutdown action

Use `chshut` and `dumpconf` to configure a CP command as a shutdown action, or directly write to the relevant sysfs attributes.

**Before you start:** This information applies to Linux on z/VM only.

In sysfs, two attributes are required to set a z/VM CP command as a shutdown action for a trigger `<trigger>`:

- `/sys/firmware/shutdown_actions/on_<trigger>` must be set to `vmcmd`.
- `/sys/firmware/vmcmd/on_<trigger>` specifies the z/VM CP command.

The values of the attributes in the `/sys/firmware/vmcmd` directory must conform to these rules:
- The value must be a valid z/VM CP command.
- The commands, including any z/VM user IDs or device numbers, must be specified with uppercase characters.
- Commands that include blanks must be delimited by double quotation marks (").
- The value must not exceed 127 characters.

You can specify multiple z/VM CP commands that are separated by the newline character "\n". Each newline is counted as one character. When writing values with multiple commands, use this syntax to ensure that the newline character is processed correctly:

```bash
# echo -e <cmd1>

<cmd2>... | cat > /sys/firmware/vmcmd/on_<trigger>
```

where `<cmd1>`\n`<cmd2>`... are two or more z/VM CP commands and `on_<trigger>` is one of the attributes in the `vmcmd` directory.

The `-e echo` option and redirect through `cat` are required because of the newline character.

**Example for a single z/VM CP command**

Issue the following command to configure the z/VM CP LOGOFF command as the shutdown action for the `poff` shutdown trigger.

```bash
# chshut poff vmcmd "LOGOFF"
```

Alternatively, you can issue the following commands to directly write the shutdown configuration to sysfs:

```bash
# echo vmcmd > /sys/firmware/shutdown_actions/on_poff
# echo LOGOFF > /sys/firmware/vmcmd/on_poff
```

Figure 32 on page 113 illustrates the relationship of the sysfs attributes for this example.
Example for multiple z/VM CP commands

Issue the following command to configure two z/VM CP commands as the shutdown action for the poff shutdown trigger. First a message is sent to user OPERATOR, and then the LOGOFF command is issued.

```bash
# chshut poff vmcmd "MSG OPERATOR Going down" vmcmd "LOGOFF"
```

Alternatively, you can issue the following commands to directly write the shutdown configuration to sysfs:

```bash
# echo vmcmd > /sys/firmware/shutdown_actions/on_poff
# echo -e "MSG OPERATOR Going down\nLOGOFF" | cat > /sys/firmware/vmcmd/on_poff
```
Chapter 9. The diag288 watchdog device driver

The watchdog device driver provides Linux watchdog applications with access to the watchdog timer.

You can use the diag288 watchdog in these environments:
- Linux on z/VM
- Linux in LPAR mode as of z13s® and z13 with the enhancements of February 2016.

The diag288 watchdog device driver provides the following features:
- Access to the watchdog timer on Z.
- An API for watchdog applications (see “External programming interfaces” on page 118).

Watchdog applications can be used to set up automated restart mechanisms. Watchdog-based restart mechanisms are an alternative to a networked heartbeat with STONITH.

Watchdog applications that communicate directly with the Z firmware or with the z/VM control program (CP) do not require a third operating system to monitor a heartbeat.

What you should know about the diag288 watchdog device driver

The watchdog function comprises two components: a watchdog application that runs on the Linux instance being controlled and a watchdog timer outside the Linux instance.

While the Linux instance operates satisfactorily, the watchdog application reports a positive status to the watchdog timer at regular intervals. The watchdog application uses a device node to pass these status reports to the timer (Figure 33).

The watchdog application typically derives its status by monitoring critical network connections, file systems, and processes on the Linux instance. If a specified time elapses without a positive report being received by the watchdog timer, the watchdog timer assumes that the Linux instance is in an error state. The watchdog timer then triggers a predefined action against the Linux instance. For example, Linux might be shut down or rebooted, or a system dump might be
initiated. For information about setting the default timer and performing other actions, see “External programming interfaces” on page 118.

Linux on z/VM only: Loading or saving a DCSS can take a long time during which the virtual machine does not respond, depending on the size of the DCSS. As a result, a watchdog might time out and restart the guest. You are advised not to use the watchdog in combination with loading or saving DCSSs.

See also the generic watchdog documentation in the Linux kernel source tree under Documentation/watchdog.

**Loading and configuring the diag288 watchdog device driver**

You configure the diag288 watchdog device driver when you load the module.

```bash
modprobe diag288_wdt cmd="SYSTEM RESTART" conceal=1
```

Notes:

1. `cmd=` and `conceal=` apply only to Linux on z/VM and are ignored for Linux in LPAR mode.

where:

- `<command>`
  - configures the shutdown action to be taken if Linux on z/VM fails. The default, “SYSTEM RESTART”, configures the shutdown action that is specified for the restart shutdown trigger (see Chapter 8, “Shutdown actions,” on page 109).
  - Any other specification dissociates the timeout action from the restart shutdown trigger. Instead, the specification is issued by CP and must adhere to these rules:
    - It must be a single valid CP command
    - It must not exceed 230 characters
    - It must be enclosed by quotation marks if it contains any blanks or newline characters
  - The specification is converted from ASCII to uppercase EBCDIC.

For details about CP commands see z/VM CP Commands and Utilities Reference, SC24-6175.

On an running instance of Linux on z/VM, you can write to `/sys/module/diag288_wdt/parameters/cmd` to replace the command you specify...
when loading the module. Through this sysfs interface, you can also specify multiple commands to be issued, see Examples for more details.

The preferred method for configuring a timeout action other than a system restart is to configure a different shutdown action for the restart shutdown trigger.

`conceal=1`

enables the protected application environment where the guest is protected from unexpectedly entering CP READ. Do not enable the protected environment for guests with multiprocessor configurations. The protected application facility supports only virtual uniprocessor systems.

For details, see the “SET CONCEAL” section of z/VM CP Commands and Utilities Reference, SC24-6175.

`<nowayout_flag>`

determines what happens when the watchdog device node is closed by the watchdog application.

If the flag is set to 1 (default), the watchdog timer keeps running and triggers an action if no positive status report is received within the specified time interval. If the character "V" is written to the device and the flag is set to 0, the z/VM watchdog timer is stopped and the Linux instance continues without the watchdog support.

Examples for Linux on z/VM

The following command loads the watchdog module and determines that, on failure, the Linux instance is to be IPLed from a device with devno 0xb1a0. The protected application environment is not enabled. The watchdog application can close the watchdog device node after writing "V" to it. As a result the watchdog timer becomes ineffective and does not IPL the guest.

```
# modprobe diag288_wdt cmd="ipl b1a0" nowayout=0
```

The following example shows how to specify multiple commands to be issued.

```
# /usr/bin/printf "<cmd1>
<cmd2>
<cmd3>" > /sys/module/diag288_wdt/parameters/cmd
```

where `<cmd1>`, `<cmd2>`, and `<cmd3>` are z/VM commands.

Use the `printf` version at `/usr/bin/printf`. The built-in `printf` command from bash might not process the newline characters as intended.

To verify that your commands have been accepted, issue: To verify that your commands have been accepted, issue:

```
# cat /sys/module/diag288_wdt/parameters/cmd
<cmd1>
<cmd2>
<cmd3>
```

Note: You cannot specify multiple commands as module parameters while loading the module.
Setting the timeout action

The timeout action for the diag288 watchdog device driver is defined by the restart shutdown trigger.

The default action is a **PSW restart** for Linux in LPAR mode and the CP system **restart** command for Linux on z/VM. You can change how Linux reacts to a PSW restart by changing the shutdown action for the restart shutdown trigger (see [Chapter 8, “Shutdown actions,” on page 109](#)).

For Linux on z/VM, you can use the diag288.cmd= kernel parameter or the cmd= module parameter to directly specify a z/VM CP command to be issued, independent of the restart shutdown trigger.

External programming interfaces

There is an API for applications that work with the watchdog device driver.

**Application programmers:** This information is intended for programmers who want to write watchdog applications that work with the watchdog device driver.

For information about the API and the supported IOCTLs, see the Documentation/watchdog/watchdog-api.txt file in the Linux source tree.

The default watchdog timeout is 30 seconds, the minimum timeout that can be set through the IOCTL WDIOC_SETTIMEOUT is 15 seconds.
# Part 3. Storage

<table>
<thead>
<tr>
<th>Chapter 10. DASD device driver</th>
<th>121</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>121</td>
</tr>
<tr>
<td>What you should know about DASD</td>
<td>122</td>
</tr>
<tr>
<td>Setting up the DASD device driver</td>
<td>131</td>
</tr>
<tr>
<td>Working with DASDs</td>
<td>134</td>
</tr>
</tbody>
</table>

| Chapter 11. SCSI-over-Fibre Channel device driver | 159 |
| Features                                      | 159 |
| What you should know about zfcp                | 159 |
| Setting up the zfcp device driver              | 165 |
| Working with FCP devices                       | 167 |
| Working with target ports                      | 174 |
| Working with SCSI devices                      | 181 |
| Confirming end-to-end data consistency checking | 194 |
| Scenario for finding available LUNs            | 195 |
| zfcp HBA API support                           | 196 |

| Chapter 12. Storage-class memory device driver | 201 |
| Features                                      | 201 |
| What you should know about storage-class memory | 201 |
| Setting up the storage-class memory device driver | 202 |
| Working with storage-class memory increments   | 202 |

| Chapter 13. Channel-attached tape device driver | 205 |
| Features                                      | 205 |
| What you should know about channel-attached tape devices | 205 |
| Loading the tape device driver                | 208 |
| Working with tape devices                     | 208 |

| Chapter 14. XPRAM device driver              | 215 |
| XPRAM features                                | 215 |
| What you should know about XPRAM.             | 215 |
| Setting up the XPRAM device driver            | 216 |

Red Hat Enterprise Linux 7.5 includes several storage device drivers that are specific to z/Architecture.

**Newest version**

You can find the newest version of this book at


**Restrictions**

For prerequisites and restrictions see the IBM Z architecture specific information in the Red Hat Enterprise Linux 7.5 release notes at

Chapter 10. DASD device driver

The DASD device driver provides access to all real or emulated direct access storage devices (DASD) that can be attached to the channel subsystem of an IBM mainframe.

DASD devices include various physical media on which data is organized in blocks or records or both. The blocks or records in a DASD can be accessed for read or write in random order.

Traditional DASD devices are attached to a control unit that is connected to a mainframe I/O channel. Today, these real DASDs have been largely replaced by emulated DASDs. For example, emulated DASDs can be the volumes of the IBM System Storage® DS8000® Turbo, or the volumes of the IBM System Storage DS6000™. These emulated DASD are completely virtual and the identity of the physical device is hidden.

SCSI disks that are attached through an FCP channel are not classified as DASD. They are handled by the zfcp driver (see Chapter 11, “SCSI-over-Fibre Channel device driver,” on page 159).

Features

The DASD device driver supports a wide range of disk devices and disk functions.

- The DASD device driver has no dependencies on the adapter hardware that is used to physically connect the DASDs to the IBM Z hardware. You can use any adapter that is supported by the IBM Z hardware (see www.ibm.com/systems/z/connectivity for more information).
- The DASD device driver supports ESS virtual ECKD type disks
- The DASD device driver supports the control unit attached physical ECKD (Extended Count Key Data) and FBA (Fixed Block Access) devices as summarized in Table 14

<table>
<thead>
<tr>
<th>Device format</th>
<th>Control unit type</th>
<th>Device type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECKD</td>
<td>1750</td>
<td>3380 and 3390</td>
</tr>
<tr>
<td>ECKD</td>
<td>2107</td>
<td>3380 and 3390</td>
</tr>
<tr>
<td>ECKD</td>
<td>2105</td>
<td>3380 and 3390</td>
</tr>
<tr>
<td>ECKD</td>
<td>3990</td>
<td>3380 and 3390</td>
</tr>
<tr>
<td>ECKD</td>
<td>9343</td>
<td>9345</td>
</tr>
<tr>
<td>ECKD</td>
<td>9380</td>
<td>3390</td>
</tr>
<tr>
<td>FBA</td>
<td>6310</td>
<td>9336</td>
</tr>
<tr>
<td>FBA</td>
<td>3880</td>
<td>3370</td>
</tr>
</tbody>
</table>

All models of the specified control units and device types can be used with the DASD device driver. This includes large devices with more than 65520 cylinders, for example, 3390 Model A. Check the storage support statement to find out what works for Red Hat Enterprise Linux.
• The DASD device driver provides a disk format with up to three partitions per disk. See “IBM Z compatible disk layout” on page 123 for details.

• The DASD device driver provides an option for extended error reporting for ECKD devices. Extended error reporting can support high availability setups.

• The DASD device driver supports parallel access volume (PAV) and HyperPAV on storage devices that provide this feature. For more information about PAV and HyperPAV, see How to Improve Performance with PAV, SC33-8414. Use the dasdstat command to check whether a DASD uses PAV, see “Scenario: Verifying that PAV and HPF are used” on page 148.

• The DASD device driver supports High Performance FICON, including multitrack requests, on storage devices that provide this feature. Use the dasdstat command to check whether a DASD uses High Performance FICON, see “Scenario: Verifying that PAV and HPF are used” on page 148.

• The DASD device driver supports large volumes (devices with more than 65520 cylinders, for example, 3390 Model A), solid state devices, and encrypted devices.

What you should know about DASD

The DASD device driver supports various disk layouts with different partitioning capabilities. The DASD device naming scheme helps you to keep track of your DASDs and DASD device nodes.

The IBM label partitioning scheme

Linux on Z supports the same standard DASD format that is also used by traditional mainframe operating systems, but it also supports any other Linux partition table. The DASD device driver is embedded into the Linux generic support for partitioned disks. As a result, you can use any partition table format that is supported by Linux for your DASDs.

Traditional mainframe operating systems (such as, z/OS, z/VM, and z/VSE®) expect a standard DASD format. In particular, the format of the first two tracks of a DASD is defined by this standard. These tracks include the Z IPL, label, and for some layouts VTOC records. Partitioning schemes for platforms other than Z generally do not preserve these mainframe specific records.

Red Hat Enterprise Linux 7.5 for z Systems includes the IBM label partitioning scheme that preserves the Z IPL, label, and VTOC records. With this partitioning scheme, Linux can share a disk with other mainframe operating systems. For example, a traditional mainframe operating system can handle backup and restore for a partition that is used by Linux.

The following sections describe the layouts that are supported by the IBM label partitioning scheme:

• “IBM Z compatible disk layout” on page 123
• “Linux disk layout” on page 126
• “CMS disk layout” on page 126

DASD partitions

Partitioning DASD has the same advantages as for other disk types, but there are some prerequisites and a special tool, fdasd.
A DASD partition is a contiguous set of DASD blocks that is treated by Linux as an independent disk and by the traditional mainframe operating systems as a data set.

With the Linux disk layout (LDL) and the CMS disk layout, you always have a single partition only. This partition is defined by the LDL or CMS formatted area of the disk. With the compatible disk layout, you can have up to three partitions.

There are several reasons why you might want to have multiple partitions on a DASD, for example:

**Limit data growth**
Runaway processes or undisciplined users can consume disk space to an extent that the operating system runs short of space for essential operations. Partitions can help to isolate the space that is available to particular processes.

**Encapsulate your data**
If a file system gets damaged, this damage is likely to be restricted to a single partition. Partitioning can reduce the scope of data damage.

**Recommendations:**
- Use `fdasd` to create or alter partitions on ECKD type DASD that are formatted with the compatible disk layout. If you use another partition editor, it is your responsibility to ensure that partitions do not overlap. If they do, data damage occurs.
- Leave no gaps between adjacent partitions to avoid wasting space. Gaps are not reported as errors, and can be reclaimed only by deleting and re-creating one or more of the surrounding partitions and rebuilding the file system on them.

A disk need not be partitioned completely. You can begin by creating only one or two partitions at the start of your disk and convert the remaining space to a partition later.

There is no facility for moving, enlarging, or reducing partitions, because `fdasd` has no control over the file system on the partition. You can only delete and re-create them. Changing the partition table results in loss of data in all altered partitions. It is up to you to preserve the data by copying it to another medium.

**IBM Z compatible disk layout**
With the compatible disk layout a DASD can have up to three partitions that can be accessed by traditional mainframe operating systems.

You can only format ECKD type DASD with the compatible disk layout.

[Figure 34](#) illustrates a DASD with the compatible disk layout.

```
<table>
<thead>
<tr>
<th>IPL records</th>
<th>VOL1</th>
<th>VTOC</th>
<th>/dev/dasd&lt;x&gt;1</th>
<th>/dev/dasd&lt;x&gt;2</th>
<th>/dev/dasd&lt;x&gt;3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>/dev/dasd&lt;x&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

[Figure 34. Compatible disk layout](#)
The IPL records, volume label (VOL1), and VTOC of disks with the compatible 
disk layout are on the first two tracks of the disks. These tracks are not intended 
for use by Linux applications. Using the tracks can result in data loss.

Linux can address the device as a whole as /dev/dasd<x>, where <x> can be one to 
four letters that identify the individual DASD (see “DASD naming scheme” on 
page 127). See “DASD device nodes” on page 128 for alternative addressing 
possibilities.

Disks with the compatible disk layout can have one to three partitions. Linux 
addresses the first partition as /dev/dasd<x>1, the second as /dev/dasd<x>2, and 
the third as /dev/dasd<x>3.

You use the dasdfmt command (see “dasdfmt - Format a DASD” on page 520) to 
format a disk with the compatible disk layout. You use the fdasd command (see 
“fdasd – Partition a DASD” on page 539) to create and modify partitions.

**Volume label**

The volume label includes information about the disk layout, the VOLSER, and a 
pointer to the VTOC.

The DASD volume label is located in the third block of the first track of the device 
(cylinder 0, track 0, block 2). This block has a 4-byte key, and an 80-byte data area 
with the following content:

**key** for disks with the compatible disk layout, contains the four EBCDIC 
characters “VOL1” to identify the block as a volume label.

**label identifier**
is identical to the key field.

**VOLSER**
is a name that you can use to identify the DASD device. A volume serial 
number (VOLSER) can be one to six EBCDIC characters. If you want to use 
VOLSERs as identifiers for your DASD, be sure to assign unique VOLSERs.

You can assign VOLSERs from Linux by using the dasdfmt or fdasd 
command. These commands enforce that VOLSERs:

- Are alphanumeric
- Are uppercase (by uppercase conversion)
- Contain no embedded blanks
- Contain no special characters other than $, #, @, and %

**Tip:** Avoid special characters altogether.

**Note:** The VOLSER values SCRTCH, PRIVAT, MIGRAT, or Lnnnnn (An “L” 
followed by five digits) are reserved for special purposes by other 
mainframe operating systems and should not be used by Linux.

These rules are more restrictive than the VOLSERs that are allowed by the 
traditional mainframe operating systems. For compatibility, Linux tolerates 
existing VOLSERs with lowercase letters and special characters other than 
$, #, @, and %. Enclose VOLSERs with special characters in single 
quotiation marks if you must specify it, for example, as a command 
parameter.
VT OC address
contains the address of a standard IBM format 4 data set control block (DSCB). The format is: cylinder (2 bytes) track (2 bytes) block (1 byte).

All other fields of the volume label contain EBCDIC space characters (code 0x40).

VT OC
Instead of a regular Linux partition table, Red Hat Enterprise Linux 7.5 for z Systems, like other mainframe operating systems, uses a Volume Table Of Contents (VT OC).

The VT OC contains pointers to the location of every data set on the volume. These data sets form the Linux partitions.

The VT OC is on the second track (cylinder 0, track 1). It contains a number of records, each written in a separate data set control block (DSCB). The number of records depends on the size of the volume:
• One DSCB that describes the VT OC itself (format 4)
• One DSCB that is required by other operating systems but is not used by Linux.
  fdasd sets it to zeros (format 5).
• For volumes with more than 65534 cylinders, 1 DSCB (format 7)
• For each partition:
  – On volumes with 65534 or less cylinders, 1 DSCB (format 1)
  – On volumes with more than 65534 cylinders, 1 format 8 and one format 9 DSCB

The key of the format 1 or format 8 DSCB contains the data set name, which identifies the partition to z/OS, z/VM, and z/VSE.

The VT OC can be displayed with standard IBM Z tools such as VM/DITTO. A Linux DASD with physical device number 0x0193, volume label “LNX001”, and three partitions might be displayed like this example:

```
--- FILE NAME --- (SORTED BY =,NAME ,) ----- EXT BEGIN-END RELTRK,
1...5...10...15...20...25...30...35...40... SQ CYL-HD CYL-HD NUMTRKS
*** VT OC EXTENT ***  0  0  1  0  1  1,1
LINUX.VLNX001.PART0001.NATIVE  0  0  2  46 11  2,700
LINUX.VLNX001.PART0002.NATIVE  0  46 12  66 11  702,300
LINUX.VLNX001.PART0003.NATIVE  0  66 12  99 14  1002,498
*** THIS VOLUME IS CURRENTLY 100 PER CENT FULL WITH 0 TRACKS AVAILABLE
```

The 1s command on Linux might list this DASD and its partitions like this example:

```
# ls -l /dev/dasda+
brw-rw---- 1 root disk 94, 0 Jan 27 09:04 /dev/dasda
brw-rw---- 1 root disk 94, 1 Jan 27 09:04 /dev/dasda1
brw-rw---- 1 root disk 94, 2 Jan 27 09:04 /dev/dasda2
brw-rw---- 1 root disk 94, 3 Jan 27 09:04 /dev/dasda3
```
where dasda represent the whole DASD and dasda1, dasda2, and dasda3 represent the individual partitions.

**Linux disk layout**

The Linux disk layout does not have a VTOC, and DASD partitions that are formatted with this layout cannot be accessed by traditional mainframe operating systems.

You can format only ECKD type DASD with the Linux disk layout. Apart from accessing the disks as ECKD devices, you can also access them using the DASD DIAG access method. See “Enabling the DASD device driver to use the DIAG access method” on page 138 for how to enable DIAG.

Figure 35 illustrates a disk with the Linux disk layout.

![Figure 35. Linux disk layout](image)

DASDs with the Linux disk layout either have an LNX1 label or are not labeled. The first records of the device are reserved for IPL records and the volume label, and are not intended for use by Linux applications. All remaining records are grouped into a single partition. You cannot have more than a single partition on a DASD that is formatted in the Linux disk layout.

Linux can address the device as a whole as /dev/dasd<x>, where <x> can be one to four letters that identify the individual DASD (see “DASD naming scheme” on page 127). Linux can access the partition as /dev/dasd<x>1.

You use the dasdfmt command (see “dasdfmt - Format a DASD” on page 520) to format a disk with the Linux disk layout.

**CMS disk layout**

The CMS disk layout applies only to Linux on z/VM. The disks are formatted with z/VM tools.

Both ECKD or FBA type DASD can have the CMS disk layout. DASD partitions that are formatted with this layout cannot be accessed by traditional mainframe operating systems. Apart from accessing the disks as ECKD or FBA devices, you can also access them using the DASD DIAG access method.

Figure 36 on page 127 illustrates two variants of the CMS disk layout.
The first variant contains IPL records, a volume label (CMS1), and a CMS data area. Linux treats DASD like this equivalent to a DASD with the Linux disk layout, where the CMS data area serves as the Linux partition.

The second variant is a CMS reserved volume. In this variant, the DASD was reserved by a CMS RESERVE fn ft fm command. In addition to the IPL records and the volume label, DASD with the CMS disk layout also have CMS metadata. The CMS reserved file serves as the Linux partition.

For both variants of the CMS disk layout, you can have only a single Linux partition. The IPL record, volume label and (where applicable) the CMS metadata, are not intended for use by Linux applications.

Addressing the device and partition is the same for both variants. Linux can address the device as a whole as `/dev/dasd<x>`, where `<x>` can be one to four letters that identify the individual DASD (see “DASD naming scheme”). Linux can access the partition as `/dev/dasd<x>1`.

“Enabling the DASD device driver to use the DIAG access method” on page 138 describes how you can enable DIAG.

**Disk layout summary**

The available disk layouts differ in their support of device formats, the DASD DIAG access method, and the maximum number of partitions.

<table>
<thead>
<tr>
<th>Disk layout</th>
<th>ECKD device format</th>
<th>FBA device format</th>
<th>DIAG access method support (z/VM only)</th>
<th>Maximum number of partitions</th>
<th>Formatting tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compatible disk layout</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>3</td>
<td>dasdfmt</td>
</tr>
<tr>
<td>Linux disk layout</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>1</td>
<td>dasdfmt</td>
</tr>
<tr>
<td>CMS (z/VM only)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
<td>z/VM tools</td>
</tr>
</tbody>
</table>

**DASD naming scheme**

The DASD naming scheme maps device names and minor numbers to whole DASDs and to partitions.
The DASD device driver uses the major number 94. For each configured device it uses four minor numbers:

- The first minor number always represents the device as a whole, including IPL, VTOC, and label records.
- The remaining three minor numbers represent the up to three partitions.

With 1,048,576 (20-bit) available minor numbers, the DASD device driver can address 262,144 devices.

The DASD device driver uses a device name of the form dasd<x> for each DASD. In the name, <x> is one to four lowercase letters. [Table 16] shows how the device names map to the available minor numbers.

<table>
<thead>
<tr>
<th>Name for device as a whole</th>
<th>Minor number for device as a whole</th>
<th>Number of devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td>From</td>
</tr>
<tr>
<td>dasda</td>
<td>dasdz</td>
<td>0</td>
</tr>
<tr>
<td>dasdaa</td>
<td>dasdzz</td>
<td>104</td>
</tr>
<tr>
<td>dasdaaa</td>
<td>dasdzzz</td>
<td>2808</td>
</tr>
<tr>
<td>dasdaaaa</td>
<td>dasdnwtl</td>
<td>73112</td>
</tr>
<tr>
<td><strong>Total number of devices:</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The DASD device driver also uses a device name for each partition. The name of the partition is the name of the device as a whole with a 1, 2, or 3 appended to identify the first, second, or third partition. The three minor numbers that follow the minor number of the device as a whole are the minor number for the first, second, and third partition.

**Examples**

- “dasda” refers to the whole of the first disk in the system and “dasda1”, “dasda2”, and “dasda3” to the three partitions. The minor number for the whole device is 0. The minor numbers of the partitions are 1, 2, and 3.
- “dasdz” refers to the whole of the 101st disk in the system and “dasdz1”, “dasdz2”, and “dasdz3” to the three partitions. The minor number for the whole device is 100. The minor numbers of the partitions are 101, 102, and 103.
- “dasdaa” refers to the whole of the 102nd disk in the system and “dasdaa1”, “dasdaa2”, and “dasdaa3” to the three partitions. The minor number for the whole device is 104. The minor numbers of the partitions are 105, 106, and 107.

**DASD device nodes**

Red Hat Enterprise Linux 7.5 uses udev to create multiple device nodes for each DASD that is online.

**Device nodes based on device names**

udev creates device nodes that match the device names used by the kernel. These standard device nodes have the form /dev/<name>.

The mapping between standard device nodes and the associated physical disk space can change, for example, when you reboot Linux. To ensure that you access the intended physical disk space, you need device nodes that are based on properties that identify a particular DASD.
udev creates additional devices nodes that are based on the following information:

- The bus ID of the disk
- The disk label (VOLSER)
- The universally unique identifier (UUID) of the file system on the disk
- If available: The label of the file system on the disk

**Device nodes based on bus IDs**

udev creates device nodes of the form

```
/dev/disk/by-path/ccw-<device_bus_id>
```

for whole DASD and

```
/dev/disk/by-path/ccw-<device_bus_id>-part<n>
```

for the `<n>`th partition.

**Device nodes that are based on VOLSERs**

udev creates device nodes of the form

```
/dev/disk/by-id/ccw-<volser>
```

for whole DASD and

```
/dev/disk/by-id/ccw-<volser>-part<n>
```

for the `<n>`th partition.

If you want to use device nodes based on VOLSER, be sure that the VOLSERs in your environment are unique (see "Volume label" on page 124).

If you assign the same VOLSER to multiple devices, Linux can still access each device through its standard device node. However, only one of the devices can be accessed through the VOLSER-based device node. Thus, the node is ambiguous and might lead to unintentional data access.

Furthermore, if the VOLSER on the device that is addressed by the node is changed, the previously hidden device is not automatically addressed instead. To reassign the node, you must reboot Linux or force the kernel to reread the partition tables from disks, for example, by issuing:

```
# blockdev --rereadpt /dev/dasdzzz
```

You can assign VOLSERs to ECKD type devices with `dasdfmt` when formatting or later with `fdasd` when creating partitions.

**Device nodes based on file system information**

udev creates device nodes of the form

```
/dev/disk/by-uuid/<uuid>
```

where `<uuid>` is the UUID for the file system in a partition.

If a file system label exists, udev also creates a node of the form

```
/dev/disk/by-label/<label>
```

There are no device nodes for the whole DASD that are based on file system information.

If you want to use device nodes that are based on file system labels, be sure that the labels in your environment are unique.
Additional device nodes

/dev/disk/by-id contains additional device nodes for the DASD and partitions, that are all based on a device identifier as contained in the uid attribute of the DASD.

Note: If you want to use device nodes that are based on file system information and VOLSER, be sure that they are unique for the scope of your Linux instance. This information can be changed by a user or it can be copied, for example when backup disks are created. If two disks with the same VOLSER or UUID are online to the same Linux instance, the matching device node can point to either of these disks.

Example

For a DASD that is assigned the device name dasdzzz, has two partitions, a device bus-ID 0.0.b100 (device number 0xb100), VOLSER LNX001, and a UUID 6dd6c43d-a792-412f-a651-0031e631caed for the first and f45e955d-741a-4cf3-86b1-380ee5177ac3 for the second partition, udev creates the following device nodes:

For the whole DASD:
- /dev/dasdzzz (standard device node according to the DASD naming scheme)
- /dev/disk/by-path/ccw-0.0.b100
- /dev/disk/by-id/ccw-LNX001

For the first partition:
- /dev/dasdzzz1 (standard device node according to the DASD naming scheme)
- /dev/disk/by-path/ccw-0.0.b100-part1
- /dev/disk/by-id/ccw-LNX001-part1
- /dev/disk/by-uuid/6dd6c43d-a792-412f-a651-0031e631caed

For the second partition:
- /dev/dasdzzz2 (standard device node according to the DASD naming scheme)
- /dev/disk/by-path/ccw-0.0.b100-part2
- /dev/disk/by-id/ccw-LNX001-part2
- /dev/disk/by-uuid/f45e955d-741a-4cf3-86b1-380ee5177ac3

Accessing DASD by udev-created device nodes

Use udev-created device nodes to access a particular physical disk space, regardless of the device name that is assigned to it.

Example

The following example is based on these assumptions:
- A DASD with bus ID 0.0.b100 has two partitions.
- The standard device node of the DASD is dasdzzz.
- udev creates the following device nodes for a DASD and its partitions:
  /dev/disk/by-path/ccw-0.0.b100
  /dev/disk/by-path/ccw-0.0.b100-part1
  /dev/disk/by-path/ccw-0.0.b100-part2

Instead of issuing:
In the file system information in /etc/fstab replace the following specifications:

/dev/dasdzzz1 /temp1 ext4 defaults 0 0
/dev/dasdzzz2 /temp2 ext4 defaults 0 0

with these specifications:

/dev/disk/by-path/ccw-0.0.b100-part1 /temp1 ext4 defaults 0 0
/dev/disk/by-path/ccw-0.0.b100-part2 /temp2 ext4 defaults 0 0

You can make similar substitutions with other device nodes that udev provides for you (see “DASD device nodes” on page 128).

### Setting up the DASD device driver

Unless the DASD device driver modules are loaded for you during the boot process, load and configure them with the `modprobe` command.

---

**DASD module parameter syntax**

```
modprobe dasd mod,

dasd=
device-spec

eer_pages=<pages>

dasd_eckd_mod

dasd_fba_mod

dasd_diag_mod

dasd=
device-spec

device-spec:

<device_bus_id>

<from_device_bus_id><to_device_bus_id>

ro

diag

erplog

failfast
```

Where:

**dasd_mod**

loads the device driver base module.

When you are loading the base module you can specify the dasd= parameter.
You can use the `eerr_pages` parameter to determine the number of pages that are used for internal buffering of error records.

**autodetect**
causes the DASD device driver to allocate device names and the corresponding minor numbers to all DASD devices and set them online during the boot process. See "DASD naming scheme" on page 127 for the naming scheme.

The device names are assigned in order of ascending subchannel numbers. Auto-detection can yield confusing results if you change your I/O configuration and reboot, or if your Linux instance runs as a z/VM guest because the devices might appear with different names and minor numbers after rebooting.

**probeonly**
causes the DASD device driver to reject any “open” syscall with EPERM.

**autodetect, probeonly**
causes the DASD device driver to assign device names and minor numbers as for auto-detect. All devices regardless of whether they are accessible as DASD return EPERM to any “open” requests.

**nopav**
suppresses parallel access volume (PAV and HyperPAV) enablement for Linux instances that run in LPAR mode. The `nopav` keyword has no effect for Linux on z/VM.

**nofc**
suppresses accessing the storage server with the I/O subsystem in transport mode (also known as High Performance FICON).

**<device_bus_id>**
specifies a single DASD.

**<from_device_bus_id>-<to_device_bus_id>**
specifies the first and last DASD in a range. All DASD devices with bus IDs in the range are selected. The device bus-IDs `<from_device_bus_id>` and `<to_device_bus_id>` need not correspond to actual DASD.

**(ro)** accesses the specified device or device range in read-only mode.

**(diag)** forces the device driver to access the device (range) with the DIAG access method.

**(erplog)** enables enhanced error recovery processing (ERP) related logging through syslogd. If `erplog` is specified for a range of devices, the logging is switched on during device initialization.

**(failfast)** immediately returns “failed” for an I/O operation when the last path to a DASD is lost.

**Attention:** Enable immediate failure of I/O requests only in setups where a failed I/O request can be recovered outside the scope of a single DASD (see "Enabling and disabling immediate failure of I/O requests" on page 142).

**dasd_eckd_mod**
loads the ECKD module.

**dasd_fba_mod**
loads the FBA module.
**dasd_diag_mod**
loads the DIAG module.

If you supply a DASD module parameter with device specifications
dasd=<device-list1>,<device-list2> ..., the device names and minor numbers
are assigned in the order in which the devices are specified. The names and
corresponding minor numbers are always assigned, even if the device is not
present, or not accessible. For information about including device specifications in
a boot configuration, see “Including parameters for modules in a boot
configuration” on page 29.

If you use **autodetect** in addition to explicit device specifications, device names
are assigned to the specified devices first and device-specific parameters, like **ro**,
are honored. The remaining devices are handled as described for **autodetect**.

The DASD base component is required by the other modules. **modprobe** takes care
of this dependency for you and ensures that the base module is loaded
automatically, if necessary.

**Hint:** **modprobe** might return before udev has created all device nodes for the
specified DASDs. If you need to assure that all nodes are present, for example in
scripts, follow the **modprobe** command with:

```
# udevadm settle
```

For command details see the **modprobe** man page.

**Example**

The following example specifies a range of DASD devices and two individual
DASD devices:

```
modprobe dasd_mod dasd=0.0.7000-0.0.7002,0.0.7005(ro),0.0.7006
```

**Table 17** shows the resulting allocation of device names:

**Table 17. Example mapping of device names to devices**

<table>
<thead>
<tr>
<th>Name</th>
<th>To access</th>
</tr>
</thead>
<tbody>
<tr>
<td>dasda</td>
<td>device 0.0.7000 as a whole</td>
</tr>
<tr>
<td>dasda1</td>
<td>the first partition on 0.0.7000</td>
</tr>
<tr>
<td>dasda2</td>
<td>the second partition on 0.0.7000</td>
</tr>
<tr>
<td>dasda3</td>
<td>the third partition on 0.0.7000</td>
</tr>
<tr>
<td>dasdb</td>
<td>device 0.0.7001 as a whole</td>
</tr>
<tr>
<td>dasdb1</td>
<td>the first partition on 0.0.7001</td>
</tr>
<tr>
<td>dasdb2</td>
<td>the second partition on 0.0.7001</td>
</tr>
<tr>
<td>dasdb3</td>
<td>the third partition on 0.0.7001</td>
</tr>
<tr>
<td>dasdc</td>
<td>device 0.0.7002 as a whole</td>
</tr>
<tr>
<td>dasdc1</td>
<td>the first partition on 0.0.7002</td>
</tr>
<tr>
<td>dasdc2</td>
<td>the second partition on 0.0.7002</td>
</tr>
<tr>
<td>dasdc3</td>
<td>the third partition on 0.0.7002</td>
</tr>
<tr>
<td>dasdd</td>
<td>device 0.0.7005 as a whole</td>
</tr>
<tr>
<td>dasdd1</td>
<td>the first partition on 0.0.7005 (read-only)</td>
</tr>
</tbody>
</table>
Table 17. Example mapping of device names to devices (continued)

<table>
<thead>
<tr>
<th>Name</th>
<th>To access</th>
</tr>
</thead>
<tbody>
<tr>
<td>dasdd2</td>
<td>the second partition on 0.7005 (read-only)</td>
</tr>
<tr>
<td>dasdd3</td>
<td>the third partition on 0.7005 (read-only)</td>
</tr>
<tr>
<td>dasde</td>
<td>device 0.7006 as a whole</td>
</tr>
<tr>
<td>dasde1</td>
<td>the first partition on 0.7006</td>
</tr>
<tr>
<td>dasde2</td>
<td>the second partition on 0.7006</td>
</tr>
<tr>
<td>dasde3</td>
<td>the third partition on 0.7006</td>
</tr>
</tbody>
</table>

The following example specifies that High Performance FICON are to be suppressed for all DASDs:

```
modprobe dasd_mod dasd=nofcx,4711-4713
```

Working with DASDs

You might have to prepare DASD for use, configure troubleshooting functions, or configure special device features for your DASDs.

Most of the following tasks involve writing to and reading from device attributes in sysfs. This method is useful on a running system where you want to make dynamic changes. For changes that persist across IPLs, use the configuration file `/etc/zipl.conf` for DASDs that are part of the root file system and `/etc/dasd.conf` for data disks. An example of how to define a DASD device persistently is in *Red Hat Enterprise Linux 7.5 Installation Guide*. For a general discussion of configuration files, see *Red Hat Enterprise Linux 7.5 System Administrator’s Guide*.

See “Working with newly available devices” on page 10 to avoid errors when you are working with devices that have become available to a running Linux instance.

- “Preparing an ECKD type DASD for use” on page 135
- “Preparing an FBA-type DASD for use” on page 136
- “Accessing DASD by force” on page 137
- “Enabling the DASD device driver to use the DIAG access method” on page 138
- “Using extended error reporting for ECKD type DASD” on page 139
- “Setting a DASD online or offline” on page 140
- “Enabling and disabling logging” on page 142
- “Enabling and disabling immediate failure of I/O requests” on page 142
- “Setting the timeout for I/O requests” on page 143
- “Working with DASD statistics in debugfs” on page 144
- “Accessing full ECKD tracks” on page 149
- “Handling lost device reservations” on page 150
- “Reading and resetting the reservation state” on page 152
- “Setting defective channel paths offline automatically” on page 153
- “Querying the HPF setting of a channel path” on page 154
- “Displaying DASD information” on page 155
Preparing an ECKD type DASD for use

Before you can use an ECKD type DASD as a disk for Linux on Z, you must format it with a suitable disk layout. You must then create a file system or define a swap space.

Before you begin

- The modules for the base component and the ECKD component of the DASD device driver must have been loaded.
- The DASD device driver must have recognized the device as an ECKD type device.
- You must know the device bus-ID for your DASD.

About this task

If you format the DASD with the compatible disk layout, you must create one, two, or three partitions. You can then use your partitions as swap areas or to create a Linux file system.

Procedure

Perform these steps to prepare the DASD:

1. Issue `lsdasd` (see “lsdasd - List DASD devices” on page 571) to find out if the device is online. If necessary, set the device online using `chccwdev` (see “chccwdev - Set CCW device attributes” on page 490).

   Example:
   
   ```
   # chccwdev -e 0.0.b100
   ```

2. Format the device with the `dasdfmt` command (see “dasdfmt - Format a DASD” on page 520 for details). The formatting process can take hours for large DASDs. If you want to use the CMS disk layout, and your DASD is already formatted with the CMS disk layout, skip this step.

   Tips:
   - Use the largest possible block size, ideally 4096; the net capacity of an ECKD DASD decreases for smaller block sizes. For example, a DASD formatted with a block size of 512 byte has only half of the net capacity of the same DASD formatted with a block size of 4096 byte.
   - For DASDs that have previously been formatted with the cd1 or ld1 disk layout, use the `dasdfmt` quick format mode.
   - Use the `-p` option to display a progress bar.

   Example: Assuming that `/dev/dasdzzz` is a valid device node for 0.0.b100:
   
   ```
   # dasdfmt -b 4096 -p /dev/dasdzzz
   ```

3. Proceed according to your chosen disk layout:
   - If you have formatted your DASD with the Linux disk layout or the CMS disk layout, skip this step and continue with step 4 on page 136. You already have one partition and cannot add further partitions on your DASD.
   - If you have formatted your DASD with the compatible disk layout use the `fdasd` command to create up to three partitions (see “fdasd – Partition a DASD” on page 539 for details).
Example: To start the partitioning tool in interactive mode for partitioning a device /dev/dasdzzz issue:

```
# fdasd /dev/dasdzzz
```

If you create three partitions for a DASD /dev/dasdzzz, the device nodes for the partitions are /dev/dasdzzz1, /dev/dasdzzz2, and /dev/dasdzzz3.

Result: fdasd creates the partitions and updates the partition table (see "VTOR" on page 125).

4. Depending on the intended use of each partition, create a file system on the partition or define it as a swap space.
   - Either, create a file system. For example, use the Linux `mkfs.ext4` command to create an ext4 file system (see the man page for details).

   Note: Do not make the block size of the file system smaller than the block size that was used for formatting the disk with the `dasdfmt` command.

   Example:

   ```
   # mkfs.ext4 -b 4096 /dev/dasdzzz1
   ```

   - Or define the partition as a swap space with the `mkswap` command (see the man page for details).

5. Mount each file system to the mount point of your choice in Linux and enable your swap partitions.

Example: To mount a file system in a partition /dev/dasdzzz1 to a mount point /mnt and to enable a swap partition /dev/dasdzzz2 issue:

```
# mount /dev/dasdzzz1 /mnt
# swapon /dev/dasdzzz2
```

Preparing an FBA-type DASD for use

Before you can use an FBA-type DASD as a disk for Linux on Z, you must create a file system or define a swap space.

Before you begin

- The modules for the base component and the FBA component of the DASD device driver must have been loaded.
- The DASD device driver must have recognized the device as an FBA device.
- You must know the device bus-ID or the device node through which the DASD can be addressed.

Procedure

Perform these steps to prepare the DASD:

1. Depending on the intended use of the partition, create a file system on it or define it as a swap space.
   - Either create a file system, for example, with the Linux `mke2fs` command (see the man page for details).

   Example:
### Accessing DASD by force

A Linux instance can encounter DASDs that are locked by another system.

Such a DASD is referred to as “externally locked” or “boxed”. The Linux instance cannot analyze a DASD while it is externally locked.

#### About this task

To check whether a DASD has been externally locked, read its availability attribute. This attribute should be “good”. If it is “boxed”, the DASD has been externally locked. Because a boxed DASD might not be recognized as DASD, it might not show up in the device driver view in sysfs. If necessary, use the device category view instead (see “Device views in sysfs” on page 11).

**CAUTION:**
Breaking an external lock can have unpredictable effects on the system that holds the lock.

#### Procedure

1. Optional: To read the availability attribute of a DASD, issue a command of this form:

   ```
   # cat /sys/bus/ccw/devices/<device_bus_id>/availability
   ```

   **Example:** This example shows that a DASD with device bus-ID 0.0.b110 (device number 0xb110) has been externally locked.

   ```
   # cat /sys/bus/ccw/devices/0.0.b110/availability
   boxed
   ```

   If the DASD is an ECKD type DASD and if you know the device bus-ID, you can break the external lock and set the device online. This means that the lock of the external system is broken with the “unconditional reserve” channel command.

2. To force a boxed DASD online, write force to the online device attribute. Issue a command of this form:
Example: To force a DASD with device number 0xb110 online issue:

```bash
# echo force > /sys/bus/ccw/devices/0.0.b110/online
```

Results

If the external lock is successfully broken or if the lock has been surrendered by the time the command is processed, the device is analyzed and set online. If it is not possible to break the external lock (for example, because of a timeout, or because it is an FBA-type DASD), the device remains in the boxed state. This command might take some time to complete.

For information about breaking the lock of a DASD that has already been analyzed see “tunedasd - Adjust low-level DASD settings” on page 625.

Enabling the DASD device driver to use the DIAG access method

Linux on z/VM can use the DIAG access method to access DASDs with the help of z/VM functions.

Before you begin

This section applies only to Linux instances and DASD for which all of the following conditions are true:

- The Linux instance runs as a z/VM guest.
- The device can be of type ECKD with either LDL or CMS disk layout, or it can be a device of type FBA.
- The module for the DIAG component must be loaded.
- The module for the component that corresponds to the DASD type (dasd_eckd_mod or dasd_fba_mod) must be loaded.
- The DASD is offline.
- The DASD does not represent a parallel access volume alias device.

About this task

You can use the DIAG access method to access both ECKD and FBA-type DASD. You use the device’s use_diag sysfs attribute to enable or switch off the DIAG access method in a system that is online. Set the use_diag attribute to 1 to enable the DIAG access method. Set the use_diag attribute to 0 to switch off the DIAG access method (this is the default).

Alternatively, you can specify diag on the command line, for example during IPL, to force the device driver to access the device (range) with the DIAG access method.
Procedure

Issue a command of this form:

```
# echo <flag> > /sys/bus/ccw/devices/<device_bus_id>/use_diag
```

where `<device_bus_id>` identifies the DASD.

If the DIAG access method is not available and you set the `use_diag` attribute to 1, you cannot set the device online (see “Setting a DASD online or offline” on page 140).

**Note:** When switching between an enabled and a disabled DIAG access method on FBA-type DASD, first reinitialize the DASD, for example, with CMS format or by overwriting any previous content. Switching without initialization might cause data-integrity problems.

For more details about DIAG, see z/VM CP Programming Services, SC24-6179.

**Example**

In this example, the DIAG access method is enabled for a DASD with device number 0xb100.

1. Ensure that the driver is loaded:

   ```
   # modprobe dasd_diag_mod
   ```

2. Identify the sysfs CCW-device directory for the device in question and change to that directory:

   ```
   # cd /sys/bus/ccw/devices/0.0.b100/
   ```

3. Ensure that the device is offline:

   ```
   # echo 0 > online
   ```

4. Enable the DIAG access method for this device by writing '1' to the `use_diag` sysfs attribute:

   ```
   # echo 1 > use_diag
   ```

5. Use the online attribute to set the device online:

   ```
   # echo 1 > online
   ```

**Using extended error reporting for ECKD type DASD**

Control the extended error reporting feature for individual ECKD type DASD through the `eerr_enabled` sysfs attribute. Use the character device of the extended error reporting module to obtain error records.

**Before you begin**

To use the extended error reporting feature, you need ECKD type DASD

**About this task**

The extended error reporting feature is turned off by default.
Procedure

To enable extended error reporting, issue a command of this form:

```
# echo 1 > /sys/bus/ccw/devices/<device_bus_id>/eer_enabled
```

where /sys/bus/ccw/devices/<device_bus_id> represents the device in sysfs.
When it is enabled on a device, a specific set of errors generates records and might have further side effects.
Disable extended error reporting, issue a command of this form:

```
# echo 0 > /sys/bus/ccw/devices/<device_bus_id>/eer_enabled
```

What to do next

You can obtain error records for all DASD for which extended error reporting is enabled from the character device of the extended error reporting module, /dev/dasd_eer. The device supports these file operations:

- **open**
  Multiple processes can open the node concurrently. Each process that opens the node has access to the records that are created from the time the node is opened. A process cannot access records that were created before the process opened the node.

- **close**
  You can close the node as usual.

- **read**
  Blocking read and non-blocking read are supported. When a record is partially read and then purged, the next read returns an I/O error -EIO.

- **poll**
  The poll operation is typically used with non-blocking read.

Setting a DASD online or offline

Use the chccwdev command or the online sysfs attribute of the device to set
DASDs online or offline.

About this task

When Linux boots, it senses your DASD. Depending on your specification for the "dasd=" parameter, it automatically sets devices online.

Procedure

Use the chccwdev command ("chccwdev - Set CCW device attributes" on page 490) to set a DASD online or offline.
Alternatively, you can write 1 to the device's online attribute to set it online or 0 to set it offline. In contrast to the sysfs attribute, the chccwdev command triggers a cio_settle for you and waits for the cio_settle to complete.
Outstanding I/O requests are canceled when you set a device offline. To wait indefinitely for outstanding I/O requests to complete before setting the device offline, use the chccwdev option --safeoffline or the sysfs attribute safe_offline.
When you set a DASD offline, the deregistration process is synchronous, unless the device is disconnected. For disconnected devices, the deregistration process is asynchronous.

**Examples**

- To set a DASD with device bus-ID 0.0.b100 online, issue:
  
  ```bash
  # chccwdev -e 0.0.b100
  
  or
  
  # echo 1 > /sys/bus/ccw/devices/0.0.b100/online
  
  • To set a DASD with device bus-ID 0.0.b100 offline, issue:
  
  ```bash
  # chccwdev -d 0.0.b100
  
  or
  
  # echo 0 > /sys/bus/ccw/devices/0.0.b100/online
  
  • To complete outstanding I/O requests and then set a DASD with device bus-ID 0.0.4711 offline, issue:
  
  ```bash
  # chccwdev -s 0.0.4711
  
  or
  
  # echo 1 > /sys/bus/ccw/devices/0.0.4711/safe_offline
  
  If an outstanding I/O request is blocked, the command might wait forever. Reasons for blocked I/O requests include reserved devices that can be released or disconnected devices that can be reconnected.

  1. Try to resolve the problem that blocks the I/O request and wait for the command to complete.

  2. If you cannot resolve the problem, issue `chccwdev -d` to cancel the outstanding I/O requests. The data is lost.

**Dynamic attach and detach**

You can dynamically attach devices to a running Red Hat Enterprise Linux 7.5 for z Systems instance, for example, from z/VM.

When a DASD is attached, Linux attempts to initialize it according to the DASD device driver configuration. You can then set the device online. You can automate setting dynamically attached devices online by using CCW hotplug events (see “CCW hotplug events” on page 18).

**Attention:** Do not detach a device that is still being used by Linux. Detaching devices might cause the system to hang or crash. Ensure that you unmount a device and set it offline before you detach it.

See “Working with newly available devices” on page 10 to avoid errors when working with devices that have become available to a running Linux instance.
Be careful to avoid errors when working with devices that have become available to a running Linux instance.

**Enabling and disabling logging**

Use the dasd= kernel or module parameter or use the erplog sysfs attribute to enable or disable error recovery processing (ERP) logging.

**Procedure**

You can enable and disable error recovery processing (ERP) logging on a running system. There are two methods:

- Use the dasd= parameter when you load the base module of the DASD device driver.

  **Example:**

  To define a device range (0.0.7000-0.0.7005) and enable, change the parameter line to contain:
  
  `dasd=0.0.7000-0.0.7005(erplog)`

- Use the sysfs attribute erplog to turn ERP-related logging on or off.

  Logging can be enabled for a specific device by writing 1 to the erplog attribute.

  **Example:**

  

  ```
  echo 1 > /sys/bus/ccw/devices/<device_bus_id>/erplog
  ```

  To disable logging, write 0 to the erplog attribute.

  **Example:**

  

  ```
  echo 0 > /sys/bus/ccw/devices/<device_bus_id>/erplog
  ```

**Enabling and disabling immediate failure of I/O requests**

Prevent devices in mirror setups from being blocked while paths are unavailable by making I/O requests fail immediately.

**About this task**

By default, if all path have been lost for a DASD, the corresponding device in Linux waits for one of the paths to recover. I/O requests are blocked while the device is waiting.

If the DASD is part of a mirror setup, this blocking might cause the entire virtual device to be blocked. You can use the failfast attribute to immediately return I/O requests as failed while no path to the device is available.

**Attention:** Use this attribute with caution and only in setups where a failed I/O request can be recovered outside the scope of a single DASD.

**Procedure**

Use one of these methods:
You can enable immediate failure of I/O requests when you load the base module of the DASD device driver.

**Example:**

To define a device range (0.0.7000-0.0.7005) and enable immediate failure of I/O requests specify:

```
dasd=0.0.7000-0.0.7005(failfast)
```

You can use the sysfs attribute `failfast` of a DASD to enable or disable immediate failure of I/O requests on or off. To enable immediate failure of I/O requests, write 1 to the `failfast` attribute.

**Example:**

```
echo 1 > /sys/bus/ccw/devices/<device_bus_id>/failfast
```

To disable immediate failure of I/O requests, write 0 to the `failfast` attribute.

**Example:**

```
echo 0 > /sys/bus/ccw/devices/<device_bus_id>/failfast
```

### Setting the timeout for I/O requests

The timeout specifies how long Linux is to wait for a response from a storage server before it considers an I/O requests failed and cancels it.

#### About this task

The default timeout for DASD I/O requests depends on the type of DASD:

- **ECKD** uses the default that is provided by the storage server.
- **FBA** 300 s
- **DIAG** 50 s

#### Procedure

You can use the `expires` attribute of a DASD to change the timeout value for that DASD.

1. To find out the current timeout value issue a command of this form:

   ```
   # cat /sys/bus/ccw/devices/<device_bus_id>/expires
   ```

2. To set the timeout to a different value issue a command of this form:

   ```
   # echo <timeout> > /sys/bus/ccw/devices/<device_bus_id>/expires
   ```

where:

- `<timeout>` is the new timeout value in seconds. The value must be an integer in the range 1 - 40,000,000.
- `<device_bus_id>` is the device bus-ID of the DASD.
Example

This example reads the current timeout value and then sets it to 120 s.

```
# cat /sys/bus/ccw/devices/0.0.7008.expires
30
# echo 120 > /sys/bus/ccw/devices/0.0.7008.expires
```

Working with DASD statistics in debugfs

Gather DASD statistics and display the data with the `dasdstat` command.

Before you begin

Instead of accessing raw DASD performance data in debugfs, you can use the `dasdstat` command to obtain more structured data (see "dasdstat - Display DASD performance statistics" on page 525).

About this task

The DASD performance data is contained in the following subdirectories of `<mountpoint>/dasd`, where `<mountpoint>` is the mount point of debugfs:

- A directory `global` that represents all available DASDs taken together.
- For each DASD, one directory with the name of the DASD block device with which the DASD is known to the DASD device driver (for example, `dasda`, `dasdb`, and `dasdc`).
- For each CCW device that corresponds to a DASD, a directory with the bus ID as the name.

Block devices that are not set up for PAV or HyperPAV map to exactly one CCW device and the corresponding directories contain the same statistics.

With PAV or HyperPAV, a bus ID can represent a base device or an alias device. Each base device is associated with a particular block device. The alias devices are not permanently associated with the same block device. At any one time, a DASD block device is associated with one or more CCW devices. Statistics that are based on bus ID, therefore, show more detail for PAV and HyperPAV setups.

Each of these directories contains a file `statistics` that you can use to perform these tasks:

- Start and stop data gathering.
- Reset statistics counters.
- Read statistics.

To control data gathering at the scope of a directory in `<mountpoint>/dasd`, issue a command of this form:

```
# echo <keyword> > <mountpoint>/dasd/<directory>/statistics
```

Where:

- `<directory>` is one of the directories in `<mountpoint>/dasd`.
- `<keyword>` specifies the action to be taken:
on to start data gathering.
off to stop data gathering.
reset to reset the statistics counters.

To read performance data, issue a command of this form:

```bash
# cat <mountpoint>/dasd/<directory>/statistics
```

**Examples for gathering and reading DASD statistics in debugfs**

Use the echo command to start and stop data gathering for individual devices or across all DASDs. Use the cat command to access the raw performance data.

The following examples assume that debugfs is mounted at /sys/kernel/debug, see “debugfs” on page xi.

- To start data gathering for summary data across all available DASDs:
  ```bash
  # echo on > /sys/kernel/debug/dasd/global/statistics
  ```

- To stop data gathering for block device dasdb:
  ```bash
  # echo off > /sys/kernel/debug/dasd/dasdb/statistics
  ```

- To reset the counters for CCW device 0.0.b301:
  ```bash
  # echo reset > /sys/kernel/debug/dasd/0.0.b301/statistics
  ```

- To read performance data for dasda, assuming that the debugfs mount point is /sys/kernel/debug, issue:
  ```bash
  # cat /sys/kernel/debug/dasd/dasda/statistics
  start_time 1293518578.885869197
  total_requests 0
  total_sectors 0
  total_pav 0
  total_bpf 0
  histogram_sectors 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_io_times 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_time_ssch_to_irq 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_time_ssch_to_irq_weighted 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_time_build_to_ssch 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_time_build_to_irq 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_time_irq_to_end 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_time_irq_to_end_weighted 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_ccw_queue_length 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  total_read_requests 0
  total_read_sectors 0
  total_read_bpf 0
  histogram_read_sectors 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_read_time 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_read_time_ssch_to_sch 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_read_time_ssch_to_irq 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_read_time_irq_to_end 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_read_time_irq_to_end_weighted 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  histogram_read_ccw_queue_length 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  ```

**Interpreting the data rows**

The raw DASD performance data in the statistics directories in debugfs is organized into labeled data rows.

This section explains the raw data in the individual data rows of the statistics. Use the dasdstat command to obtain more structured data.
**start_time**

is the UNIX epoch time stamp when data gathering was started or when the counters were last reset.

**Tip:** Use the `date` tool to convert the time stamp to a more readily human-readable format. See the `date` man page for details.

**Single counters**

have a single integer as the statistics data. All rows with labels that begin with `total_` are of this data type.

The following rows show data for the sum of all requests, read and write:

**total_requests**

is the number of requests that have been processed.

**total_sectors**

is the sum of the sizes of all requests, in units of 512-byte sectors.

**total_pav**

is the number of requests that were processed through a PAV alias device.

**total_hpf**

is the number of requests that used High Performance FICON.

The following rows show data for read requests only:

**total_read_requests**

is the number of read requests that have been processed.

**total_read_sectors**

is the sum of the sizes of all read requests, in units of 512-byte sectors.

**total_read_pav**

is the number of read requests that were processed through a PAV alias device.

**total_read_hpf**

is the number of read requests that used High Performance FICON.

**Linear histograms**

have a series of 32 integers as the statistics data. The integers represent a histogram, with a linear scale, of the number of requests in the request queue each time a request has been queued. The first integer shows how often the request queue contained zero requests, the second integer shows how often the queue contained one request, and the n-th value shows how often the queue contained n-1 requests.

**histogram_ccw_queue_length**

is the histogram data for all requests, read and write.

**histogram_read_ccw_queue_length**

is the histogram data for read requests only.

**Logarithmic histograms**

have a series of 32 integers as the statistics data. The integers represent a histogram with a logarithmic scale:

- The first integer always represents all measures of fewer than 4 units
- The second integer represents measures of 4 or more but less than 8 units
The third integer represents measures of 8 or more but less than 16 units.

The n-th integer (1 < n < 32) represents measures of $2^n$ or more but less than $2^{n+1}$ units.

The 32nd integer represents measures of $2^{32} = 4G = 4,294,967,296$ units or more.

The following rows show data for the sum of all requests, read and write:

**histogram_sectors**

is the histogram data for request sizes. A unit is a 512-byte sector.

**histogram_io_times**

is the histogram data for the total time that is needed from creating the cqr to its completion in the DASD device driver and return to the block layer. A unit is a microsecond.

**histogram_io_times_weighted**

is the histogram data of the total time, as measured for histogram_io_times, devided by the requests size in sectors. A unit is a microsecond per sector.

This metric is deprecated and there is no corresponding histogram data for read requests.

**histogram_time_build_to_ssch**

is the histogram data of the time that is needed from creating the cqr to submitting the request to the subchannel. A unit is a microsecond.

**histogram_time_ssch_to_irq**

is the histogram data of the time that is needed from submitting the request to the subchannel until an interrupt indicates that the request has been completed. A unit is a microsecond.

**histogram_time_ssch_to_irq_weighted**

is the histogram data of the time that is needed from submitting the request to the subchannel until an interrupt indicates that the request has been completed, divided by the request size in 512-byte sectors. A unit is a microsecond per sector.

This metric is deprecated and there is no corresponding histogram data for read requests.

**histogram_time_irq_to_end**

is the histogram data of the time that is needed from return of the request from the channel subsystem, until the request is returned to the block layer. A unit is a microsecond.

The following rows show data for read requests only:

**histogram_read_sectors**

is the histogram data for read request sizes. A unit is a 512-byte sector.

**histogram_read_io_times**

is the histogram data, for read requests, for the total time that is needed from creating the cqr to its completion in the DASD device driver and return to the block layer. A unit is a microsecond.
**Scenario: Verifying that PAV and HPF are used**

Use the `dasdstat` command to display DASD performance statistics, including statistics about Parallel Access Volume (PAV) and High Performance FICON (HPF).

**Procedure**

1. Enable DASD statistics for the device of interest.

   **Example:**
   ```
   # dasdstat -e dasdc
   enable statistic "/sys/kernel/debug/dasd/dasdc/statistics"
   ```

2. Assure that I/O requests are directed to the device.

   **Hints:**
   - Access a partition, rather than the whole device, to avoid directing the I/O request towards the first 2 tracks of a CDL formatted DASD. Requests to the first 2 tracks of a CDL formatted DASD are exceptional in that they never use High Performance FICON.
   - Assure that a significant I/O load is applied to the device. PAV aliases are used only if multiple I/O requests for the device are processed simultaneously.

   **Example:**
   ```
   # dd if=/dev/dasdc1 of=/dev/null bs=4k count=256
   ```

3. Look for PAV and HPF in the statistics.

   **Example:**
   ```
   # dasdstat dasdc
   -----------------------------------------------
   statistics data for statistic: dasdc
   start time of data collection: Fri Dec 11 14:22:18 CET 2015
   7 dasd I/O requests
   with 4000 sectors(512B each)
   3 requests used a PAV alias device
   7 requests used HPF
   ```
In the example, dasdc uses both Parallel Access Volume and High Performance FICON.

**Accessing full ECKD tracks**

In raw-track access mode, the DASD device driver accesses full ECKD tracks, including record zero and the count and key data fields.

**Before you begin**

- This section applies to ECKD type DASD only.
- The DASD must be offline when you change the access mode.
- The DIAG access method must not be enabled for the device.

**About this task**

With this mode, Linux can access an ECKD device regardless of the track layout. In particular, the device does not need to be formatted for Linux.

For example, with raw-track access mode Linux can create a backup copy of any ECKD device. Full-track access can also enable a special program that runs on Linux to access and process data on an ECKD device that is not formatted for Linux.

By default, the DASD device driver accesses only the data fields of ECKD devices. In default access mode, you can work with partitions, file systems, and files in the file systems on the DASD.

When using a DASD in raw-track access mode be aware that:

- In memory, each track is represented by 64 KB of data, even if the track occupies less physical disk space. Therefore, a disk in raw-track access mode appears bigger than in default mode.
- Programs must write and should read data in multiples of complete 64 KB tracks. The minimum is a single track. The maximum is eight tracks by default but can be extended to up to 16 tracks.
- The maximum number of tracks depends on the maximum number of sectors as specified in the max_sectors_kb sysfs attribute of the DASD. This attribute is located in the block device branch of sysfs at /sys/block/dasd<x>/queue/max_sectors_kb. In the path, dasd<x> is the device name assigned by the DASD device driver.
- To extend the maximum beyond eight tracks, set the max_sectors_kb to the maximum amount of data to be processed in a single read or write operation. For example, to extend the maximum to reading or writing 16 tracks at a time, set max_sectors_kb to 1024 (16 x 64).
- Programs must write only valid ECKD tracks of 64 KB.
- Programs must use direct I/O to prevent the Linux block layer from splitting tracks into fragments. Open the block device with option O_DIRECT or work with programs that use direct I/O.
- For example, the options iflag=direct and oflag=direct cause dd to use direct I/O. When using dd, also specify the block size with the bs= option. The block size determines the number of tracks that are processed in a single I/O operation. The block size must be a multiple of 64 KB and can be up to 1024 KB. Specifying a larger block size often results in better performance. If you receive disk image data from a pipe, also use the option iflag=fullblock to ensure that full blocks are written to the DASD device.
Tools cannot directly work with partitions, file systems, or files within a file system. For example, *fdasd* and *dasdfmt* cannot be used.

**Procedure**

To change the access mode, issue a command of this form:

```shell
# echo <switch> > /sys/bus/ccw/devices/<device_bus_id>/raw_track_access
```

where:

*<switch>*

is 1 to activate raw data access and 0 to deactivate raw data access.

*<device_bus_id>*

identifies the DASD.

**Example**

The following example creates a backup of a DASD 0.0.7009 on a DASD 0.0.70a1.

The initial commands ensure that both devices are offline and that the DIAG access method is not enabled for either of them. The subsequent commands activate the raw-track access mode for the two devices and set them both online. The *lsdasd* command that follows shows the mapping between device bus-IDs and device names.

The *dd* command for the copy operation specifies direct I/O for both the input and output device and the block size of 1024 KB. After the copy operation is completed, both devices are set offline. The access mode for the original device is then set back to the default and the device is set back online.

```shell
# cat /sys/bus/ccw/devices/0.0.7009/online
1
# chccwdev -d 0.0.7009
# cat /sys/bus/ccw/devices/0.0.7009/use_diag
0
# cat /sys/bus/ccw/devices/0.0.70a1/online
0
# cat /sys/bus/ccw/devices/0.0.70a1/use_diag
0
# echo 1 > /sys/bus/ccw/devices/0.0.7009/raw_track_access
# echo 1 > /sys/bus/ccw/devices/0.0.70a1/raw_track_access
# chccwdev -e 0.0.7009,0.0.70a1
# lsdasd 0.0.7009 0.0.70a1

<table>
<thead>
<tr>
<th>Bus-ID</th>
<th>Status</th>
<th>Name</th>
<th>Device Type</th>
<th>BlkSz</th>
<th>Size</th>
<th>Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.7009</td>
<td>active</td>
<td>dasdf</td>
<td>ECKD</td>
<td>4096</td>
<td>7043MB</td>
<td>1803060</td>
</tr>
<tr>
<td>0.0.70a1</td>
<td>active</td>
<td>dasdj</td>
<td>ECKD</td>
<td>4096</td>
<td>7043MB</td>
<td>1803060</td>
</tr>
</tbody>
</table>
# echo 1024 > /sys/block/dasdf/queue/max_sectors_kb
# echo 1024 > /sys/block/dasdj/queue/max_sectors_kb
# dd if=/dev/dasdf of=/dev/dasdj bs=1024k iflag=direct oflag=direct
# chccwdev -d 0.7009,0.70a1
# echo 0 > /sys/bus/ccw/devices/0.0.7009/raw_track_access
# chccwdev -e 0.0.7009
```

**Handling lost device reservations**

A DASD reservation by your Linux instance can be lost if another system unconditionally reserves this DASD.
About this task

This other system then has exclusive I/O access to the DASD for the duration of the unconditional reservation. Such unconditional reservations can be useful for handling error situations where:

- Your Linux instance cannot gracefully release the DASD.
- Another system requires access to the DASD, for example, to perform recovery actions.

After the DASD is released by the other system, your Linux instance might process pending I/O requests and write faulty data to the DASD. How to prevent pending I/O requests from being processed depends on the reservation policy. There are two reservation policies:

ignore  All I/O operations for the DASD are blocked until the DASD is released by the second system. When using this policy, reboot your Linux instance before the other system releases the DASD. This policy is the default.

fail    All I/O operations are returned as failed until the DASD is set offline or until the reservation state is reset. When using this policy, set the DASD offline and back online after the problem is resolved. See “Reading and resetting the reservation state” on page 152 about resetting the reservation state to resume operations.

Procedure

Set the reservation policy with a command of this form:

```
# echo <policy> > /sys/bus/ccw/devices/<device_bus_id>/reservation_policy
```

where:

<device_bus_id>
    specifies the DASD.

<policy>
    is one of the available policies, ignore or fail.

Examples

- The command of this example sets the reservation policy for a DASD with bus ID 0.0.7009 to fail.

  ```
  # echo fail > /sys/bus/ccw/devices/0.0.7009/reservation_policy
  ```

- This example shows a small scenario. The first two commands confirm that the reservation policy of the DASD is fail and that the reservation has been lost to another system. Assuming that the error that had occurred has already been resolved and that the other system has released the DASD, operations with the DASD are resumed by setting it offline and back online.

  ```
  # cat /sys/bus/ccw/devices/0.0.7009/reservation_policy
  fail
  # cat /sys/bus/ccw/devices/0.0.7009/last_known_reservation_state
  lost
  # chccwdev -d 0.0.7009
  # chccwdev -e 0.0.7009
  ```
Reading and resetting the reservation state

How the DASD device driver handles I/O requests depends on the last_known_reservation_state sysfs attribute of the DASD.

About this task

The last_known_reservation_state attribute reflects the reservation state as held by the DASD device driver and can differ from the actual reservation state. Use the tunedasd -Q command to find out the actual reservation state. The last_known_reservation_state sysfs attribute can have the following values:

none  The DASD device driver has no information about the device reservation state. I/O requests are processed as usual. If the DASD is reserved by another system, the I/O requests remain in the queue until they time out, or until the reservation is released.

reserved  The DASD device driver holds a valid reservation for the DASD and I/O requests are processed as usual. The DASD device driver changes this state if notified that the DASD is no longer reserved to this system. The new state depends on the reservation policy (see “Handling lost device reservations” on page 150).

ignore  The state is changed to none.

fail  The state is changed to lost.

lost  The DASD device driver had reserved the DASD, but subsequently another system has unconditionally reserved the DASD (see “Handling lost device reservations” on page 150). The device driver processes only requests that query the actual device reservation state. All other I/O requests for the device are returned as failed.

When the error that led another system to unconditionally reserve the DASD is resolved and the DASD has been released by this other system there are two methods for resuming operations:

• Setting the DASD offline and back online.
• Resetting the reservation state of the DASD.

Attention: Do not resume operations by resetting the reservation state unless your system setup maintains data integrity on the DASD despite:
• The I/O errors that are caused by the unconditional reservation
• Any changes to the DASD through the other system

You reset the reservation state by writing reset to the last_known_reservation_state sysfs attribute of the DASD. Resetting is possible only for the fail reservation policy (see “Handling lost device reservations” on page 150) and only while the value of the last_known_reservation_state attribute is lost.

To find out the reservation state of a DASD issue a command of this form:

```
# cat /sys/bus/ccw/devices/<device_bus_id>/last_known_reservation_state
```

where <device_bus_id> specifies the DASD.
Example

The command in this example queries the reservation state of a DASD with bus ID 0.0.7009.

```
# cat /sys/bus/ccw/devices/0.0.7009/last_known_reservation_state
reserved
```

Setting defective channel paths offline automatically

Control the removal of a defective channel path through the path_threshold and path_interval sysfs attributes. If a channel path does not work correctly, it is removed from normal operation if other channel paths are available.

About this task

A channel control check (CCC) is caused by any machine malfunction that affects channel-subsystem controls. An interface control check (IFCC) indicates that an incorrect signal occurred on the channel path. Usually, these errors can be recovered automatically. However, if IFCC or CCC errors occur frequently on a particular channel path, these errors indicate a failure of this channel path. Such a failure leads to performance degradation due to error recovery processing. If other channel paths are available, it might help the overall device performance to exclude the malfunctioning channel path from I/O.

The channel-path error recovery feature applies to devices for which multiple channel paths are operational. By default, the error threshold is 256 and the reset interval is 300 s (5 minutes). Accordingly, a channel path is set offline when the error count has reached 256. If 300 seconds elapse without an error the error count is reset to 0.

You can set different values through the path_threshold and path_interval sysfs attributes of the device.

Procedure

To exclude a channel path from I/O after a certain number of IFCC or CCC errors within a certain time frame, specify both path_threshold and path_interval.

1. To specify the number of errors that must occur before the channel path is taken offline, issue a command of this form:

   ```
   # echo <no_of_errors> > /sys/bus/ccw/devices/<device_bus_id>/path_threshold
   ```

   where /sys/bus/ccw/devices/<device_bus_id> represents the device in sysfs.

2. To specify the time that must elapse without errors for the counter to be reset, issue a command of this form:

   ```
   # echo <time> > /sys/bus/ccw/devices/<device_bus_id>/path_interval
   ```

Example

Setting 512 for threshold and 5 minutes (300s) for interval:

```
echo 512 > /sys/bus/ccw/devices/0.0.4711/path_threshold
echo 300 > /sys/bus/ccw/devices/0.0.4711/path_interval
```
This example leads to a deactivation of the channel path after 512 IFCCs or CCCs. When 5 minutes (300s) have passed without IFCCs or CCCs after the last error and the path was not disabled, the counter is reset.

**What to do next**

After you repair the faulty channel path, set it online again by using the `tunefs`d command with the `-p` option. See "tunefs - Adjust low-level DASD settings" on page 625 for details.

**Querying the HPF setting of a channel path**

Query the High Performance Ficon (HPF) state of a channel path through the `hpf` sysfs attribute. The HPF function can be lost if the device cannot provide the function, or if the channel path is not able to do HPF.

**About this task**

The HPF channel-path is deactivated if an HPF error occurs indicating that HPF is not available if there are other channel paths available. If no other channel paths are available, the path remains operational with HPF deactivated.

If the device loses HPF functionality, HPF is disabled for all channel paths defined for the device.

**Procedure**

High Performance FICON for a device is available if the `hpf` sysfs attribute is 1, and unavailable otherwise.

To query the HPF function for a channel path, issue a command of this form:

```
# lsdfs -1 <device_bus_id>
```

Alternatively, you can query the sysfs attribute directly:

```
# cat /sys/bus/ccw/devices/<device_bus_id>/hpf
```

where `/sys/bus/ccw/devices/<device_bus_id>` represents the device in sysfs.

**Example**

To query the availability of HPF for a device 0.0.471, issue:

```
lsdfs -1 0.0.471
```

```
0.0.471/dasdc/94:8
status: active
type: ECKD
...
hpf: 1
```

This example indicates that HPF is enabled for the device.

Alternatively, read from the `hpf` sysfs attribute:
This example indicates that HPF is disabled for device 0.4712.

**What to do next**

You can now reset the paths to the device. You can use the `tunedasd` command to reset all or one channel path.

To re-validate all paths for one device and if possible reset HPF:

```
# tunedasd --path_reset_all /dev/dasdc
Resetting all chpids for device </dev/dasdc>...
Done.
```

See “tunedasd - Adjust low-level DASD settings” on page 625 for details.

You can also use sysfs to reset a path. sysfs expects a path mask. For example to reset CHPID 44, you can use `tunedasd`:

```
tunedasd -p 44 /dev/dasde
```

This would be the same as specifying the following in sysfs:

```
echo 08 > /sys/bus/ccw/devices/0.0.9330/path_reset
```

Both commands will reset CHPID 44 (path mask 08).

**Displaying DASD information**

Use tools to display information about your DASDs, or read the attributes of the devices in sysfs.

**About this task**

There are several methods to display DASD information:

- Use `lsdasd -l` (see “Lsdasd - List DASD devices” on page 571) to display summary information about the device settings and the device geometry of multiple DASDs.
- Use `dasdview` (see “dasdview - Display DASD structure” on page 528) to display details about the contents of a particular DASD.
- Read information about a particular DASD from sysfs, as described in this section.

The sysfs representation of a DASD is a directory of the form `/sys/bus/ccw/devices/<device_bus_id>`, where `<device_bus_id>` is the bus ID of the DASD. This sysfs directory contains a number of attributes with information about the DASD.
### Table 18. Attributes with DASD information

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>alias</td>
<td>1 if the DASD is a parallel access volume (PAV) alias device. 0 if the DASD is a PAV base device or has not been set up as a PAV device.</td>
</tr>
<tr>
<td>discipline</td>
<td>Indicates the base discipline, ECKD or FBA, that is used to access the DASD. If DIAG is enabled, this attribute might read DIAG instead of the base discipline. This attribute is read-only.</td>
</tr>
<tr>
<td>eer_enabled</td>
<td>1 if the DASD is enabled for extended error reporting, 0 if it is not enabled (see “Using extended error reporting for ECKD type DASD” on page 139).</td>
</tr>
<tr>
<td>erplog</td>
<td>1 if error recovery processing (ERP) logging is enabled, 0 if ERP logging is not enabled (see “Enabling and disabling logging” on page 142).</td>
</tr>
<tr>
<td>expires</td>
<td>Indicates the time, in seconds, that Linux waits for a response to an I/O request for the DASD. If this time expires, Linux considers a request as failed and cancels it (see “Setting the timeout for I/O requests” on page 143).</td>
</tr>
<tr>
<td>failfast</td>
<td>1 if I/O operations are returned as failed immediately when the last path to the DASD is lost. 0 if a wait period for a path to return expires before an I/O operation is returned as failed. (see “Enabling and disabling immediate failure of I/O requests” on page 142).</td>
</tr>
<tr>
<td>online</td>
<td>1 if the DASD is online, 0 if it is offline (see “Setting a DASD online or offline” on page 140).</td>
</tr>
<tr>
<td>raw_track_access</td>
<td>1 if the DASD is in raw-track access mode, 0 if it is in default access mode (see “Accessing full ECKD tracks” on page 149).</td>
</tr>
<tr>
<td>readonly</td>
<td>1 if the DASD is read-only, 0 if it can be written to. This attribute is a device driver setting and does not reflect any restrictions that are imposed by the device itself. This attribute is ignored for PAV alias devices.</td>
</tr>
</tbody>
</table>
Table 18. Attributes with DASD information (continued)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>status</td>
<td>Reflects the internal state of a DASD device. Values can be:</td>
</tr>
<tr>
<td></td>
<td><strong>unknown</strong></td>
</tr>
<tr>
<td></td>
<td>Device detection has not started yet.</td>
</tr>
<tr>
<td></td>
<td><strong>new</strong></td>
</tr>
<tr>
<td></td>
<td>Detection of basic device attributes is in progress.</td>
</tr>
<tr>
<td></td>
<td><strong>detected</strong></td>
</tr>
<tr>
<td></td>
<td>Detection of basic device attributes has finished.</td>
</tr>
<tr>
<td></td>
<td><strong>basic</strong></td>
</tr>
<tr>
<td></td>
<td>The device is ready for detecting the disk layout. Low-level tools can set a device to this state when changing the disk layout, for example, when formatting the device.</td>
</tr>
<tr>
<td></td>
<td><strong>unformatted</strong></td>
</tr>
<tr>
<td></td>
<td>The disk layout detection found no valid disk layout. The device is ready for use with low-level tools like <code>dasdfmt</code>.</td>
</tr>
<tr>
<td></td>
<td><strong>ready</strong></td>
</tr>
<tr>
<td></td>
<td>The device is in an intermediate state.</td>
</tr>
<tr>
<td></td>
<td><strong>online</strong></td>
</tr>
<tr>
<td></td>
<td>The device is ready for use.</td>
</tr>
<tr>
<td>uid</td>
<td>A device identifier of the form <code>&lt;vendor&gt;..&lt;serial&gt;..&lt;subsystem_id&gt;..&lt;unit_address&gt;..&lt;minidisk_identifier&gt;</code> where</td>
</tr>
<tr>
<td></td>
<td><code>&lt;vendor&gt;</code></td>
</tr>
<tr>
<td></td>
<td>is the specification from the vendor attribute.</td>
</tr>
<tr>
<td></td>
<td><code>&lt;serial&gt;</code></td>
</tr>
<tr>
<td></td>
<td>is the serial number of the storage system.</td>
</tr>
<tr>
<td></td>
<td><code>&lt;subsystem_id&gt;</code></td>
</tr>
<tr>
<td></td>
<td>is the ID of the logical subsystem to which the DASD belongs on the storage system.</td>
</tr>
<tr>
<td></td>
<td><code>&lt;unit_address&gt;</code></td>
</tr>
<tr>
<td></td>
<td>is the address that is used within the storage system to identify the DASD.</td>
</tr>
<tr>
<td></td>
<td><code>&lt;minidisk_identifier&gt;</code></td>
</tr>
<tr>
<td></td>
<td>is an identifier that the z/VM system assigns to distinguish between minidisks on the DASD. This part of the uid is only present for Linux on z/VM and if the z/VM version and service level support this identifier.</td>
</tr>
<tr>
<td></td>
<td>This attribute is read-only.</td>
</tr>
<tr>
<td>use_diag</td>
<td>1 if the DIAG access method is enabled, 0 if the DIAG access method is not enabled (see “Enabling the DASD device driver to use the DIAG access method” on page 138). Do not enable the DIAG access method for PAV alias devices.</td>
</tr>
<tr>
<td>vendor</td>
<td>Identifies the manufacturer of the storage system that contains the DASD.</td>
</tr>
<tr>
<td></td>
<td>This attribute is read-only.</td>
</tr>
</tbody>
</table>

There are some more attributes that are common to all CCW devices (see “Device directories” on page 8).
Procedure

Issue a command of this form to read an attribute:

```
# cat /sys/bus/ccw/devices/<device_bus_id>/<attribute>
```

where `<attribute>` is one of the attributes of Table 18 on page 156.

Example

The following sequence of commands reads the attributes for a DASD with a device bus-ID 0.0.b100:

```
# cat /sys/bus/ccw/devices/0.0.b100/alias
0
# cat /sys/bus/ccw/devices/0.0.b100/discipline
ECKD
# cat /sys/bus/ccw/devices/0.0.b100/eer_enabled
0
# cat /sys/bus/ccw/devices/0.0.b100/erplog
0
# cat /sys/bus/ccw/devices/0.0.b100/expires
30
# cat /sys/bus/ccw/devices/0.0.b100/failfast
0
# cat /sys/bus/ccw/devices/0.0.b100/online
1
# cat /sys/bus/ccw/devices/0.0.b100/raw_track_access
0
# cat /sys/bus/ccw/devices/0.0.b100/readonly
1
# cat /sys/bus/ccw/devices/0.0.b100/status
online
# cat /sys/bus/ccw/devices/0.0.b100/uid
IBM,75000000092461.e900.8a
# cat /sys/bus/ccw/devices/0.0.b100/use_diag
1
# cat /sys/bus/ccw/devices/0.0.b100/vendor
IBM
```
Chapter 11. SCSI-over-Fibre Channel device driver

The SCSI-over-Fibre Channel device driver for Linux on Z (zfcp device driver) supports virtual QDIO-based SCSI-over-Fibre Channel adapters (FCP devices) and attached SCSI devices (LUNs).

Z adapter hardware typically provides multiple channels, with one port each. You can configure a channel to use the Fibre Channel Protocol (FCP). This FCP channel is then virtualized into multiple FCP devices. Thus, an FCP device is a virtual QDIO-based SCSI-over-Fibre Channel adapter with a single port.

A single physical port supports multiple FCP devices. Using N_Port ID virtualization (NPIV) you can define virtual ports and establish a one-to-one mapping between your FCP devices and virtual ports (see “N_Port ID Virtualization for FCP channels” on page 164).

On Linux, an FCP device is represented by a CCW device that is listed under /sys/bus/ccw/drivers/zfcp. Do not confuse FCP devices with SCSI devices. A SCSI device is identified by a LUN.

Features

The zfcp device driver supports a wide range of SCSI devices, various hardware adapters, specific topologies, and specific features that depend on the Z hardware.

- Linux on Z can use various SAN-attached SCSI device types, including SCSI disks, tapes, CD-ROMs, and DVDs. For a list of supported SCSI devices, see www.ibm.com/systems/z/connectivity
- SAN access through the following FCP adapters:
  - FICON Express16S+ (as of z14)
  - FICON Express16S (as of z13)
  - FICON Express8S
  - FICON Express8
  - FICON Express4

You can order hardware adapters as features for mainframe systems.

See Fibre Channel Protocol for Linux and z/VM on IBM System z, SG24-7266 for more details about using FCP with Linux on Z.

- The zfcp device driver supports switched fabric and point-to-point topologies.
- As of zEnterprise, the zfcp device driver supports end-to-end data consistency checking.
- As of FICON Express8S, the zfcp device driver supports the data router hardware feature to improve performance by reducing the path length.

For information about SCSI-3, the Fibre Channel Protocol, and Fibre Channel related information, see www.t10.org and www.t11.org.

What you should know about zfcp

The zfcp device driver is a low-level driver or host-bus adapter driver that supplements the Linux SCSI stack.
sysfs structures for FCP devices and SCSI devices

FCP devices are CCW devices. In the sysfs device driver view, remote target ports with their LUNs are nested below the FCP devices.

When Linux is booted, it senses the available FCP devices and creates directories of the form:

```
/sys/bus/ccw/drivers/zfcp/<device_bus_id>
```

where `<device_bus_id>` is the device bus-ID that corresponds to an FCP device. You use the attributes in this directory to work with the FCP device.

**Example:** `/sys/bus/ccw/drivers/zfcp/0.0.3d0c`

The zfcp device driver automatically adds port information when the FCP device is set online and when remote storage ports (`target ports`) are added. Each added target port extends this structure with a directory of the form:

```
/sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>
```

where `<wwpn>` is the worldwide port name (WWPN) of the target port. You use the attributes of this directory to work with the port.

**Example:** `/sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562`

With NPIV-enabled FCP devices, Red Hat Enterprise Linux 7.5 uses automatic LUN scanning by default. The zfcp sysfs branch ends with the target port entries. FCP devices that are not NPIV-enabled, or if automatic LUN scanning is disabled, can be configured manually, see “Configuring SCSI devices” on page 182.
Information about zfcp objects and their associated objects in the SCSI stack is distributed over the sysfs tree. To ease the burden of collecting information about zfcp devices, ports, units, and their associated SCSI stack objects, a command that is called `lszfcp` is provided with the s390utils RPM. See “lszfcp - List zfcp devices” on page 595 for more details about the command.

See also “Mapping the representations of a SCSI device in sysfs” on page 184.

**SCSI device nodes**

User space programs access SCSI devices through device nodes.

SCSI device names are assigned in the order in which the devices are detected. In a typical SAN environment, this can mean a seemingly arbitrary mapping of names to actual devices that can change between boots. Therefore, using standard device nodes of the form `/dev/<device_name>` where `<device_name>` is the device name that the SCSI stack assigns to a device, can be a challenge.

Red Hat Enterprise Linux 7.5 provides udev to create device nodes for you. Use the device nodes to identify the corresponding actual device.

**Device nodes that are based on device names**

udev creates device nodes that match the device names used by the kernel. These standard device nodes have the form `/dev/<name>`.

The examples in this section use standard device nodes as assigned by the SCSI stack. These nodes have the form `/dev/sd<x>` for entire disks and `/dev/sd<x><n>` for partitions. In these node names `<x>` represents one or more letters and `<n>` is an integer. See Documentation/devices.txt in the Linux source tree for more information about the SCSI device naming scheme.

To help you identify a particular device, udev creates device nodes that are based on the device's bus ID, the device label, and information about the file system on the device. The file system information can be a universally unique identifier (UUID) and, if available, the file system label.

**Device nodes that are based on bus IDs**

udev creates device nodes of the form

```
/dev/disk/by-path/ccw-<device_bus_id>-fc-<wwpn>-lun-<lun>
```

for whole SCSI device and

```
/dev/disk/by-path/ccw-<device_bus_id>-fc-<wwpn>-lun-<lun>-part<n>
```

for the `<n>`th partition, where `<wwpn>` is the worldwide port number of the target port and `<lun>` is the logical unit number that represents the target SCSI device.

**Note:** The format of these udev-created device nodes has changed and now matches the common code format. Device nodes of the prior form, `ccw-<device_bus_id>-zfcp-<wwpn>:<lun>` or `ccw-<device_bus_id>-zfcp-<wwpn>:<lun>-part<n>`, are also created for compatibility reasons.

**Device nodes that are based on file system information**

udev creates device nodes of the form

```
/dev/disk/by-uuid/<uuid>
```
where `<uuid>` is a unique file-system identifier (UUID) for the file system in a partition.

If a file system label has been assigned, udev also creates a node of the form

`/dev/disk/by-label/<label>`

There are no device nodes for the whole SCSI device that are based on file system information.

**Additional device nodes**

`/dev/disk/by-id` contains additional device nodes for the SCSI device and partitions, that are all based on a unique SCSI identifier generated by querying the device.

**Example**

For a SCSI device that is assigned the device name sda, has two partitions labeled boot and SWAP-sda2, a device bus-ID 0.0.3c1b (device number 0x3c1b), and a UUID 7eaf9c95-55ac-4e5e-8f18-065b313e63ca for the first and b4a818c8-747c-40a2-bfa2-acaa3ef70ead for the second partition, udev creates the following device nodes:

For the whole SCSI device:
- `/dev/sda` (standard device node according to the SCSI device naming scheme)
- `/dev/disk/by-path/ccw-0.0.3c1b-fc-0x500507630300c562-lun-0x401040ea00000000`
- `/dev/disk/by-id/scsi-36005076303ffc562000000000000010ea`
- `/dev/disk/by-id/wwn-0x6005076303ffc562000000000000010ea`

For the first partition:
- `/dev/sda1` (standard device node according to the SCSI device naming scheme)
- `/dev/disk/by-path/ccw-0.0.3c1b-fc-0x500507630300c562-lun-0x401040ea00000000-part1`
- `/dev/disk/by-uuid/7eaf9c95-55ac-4e5e-8f18-065b313e63ca`
- `/dev/disk/by-label/boot`
- `/dev/disk/by-id/scsi-36005076303ffc562000000000000010ea-part1`
- `/dev/disk/by-id/wwn-0x6005076303ffc562000000000000010ea-part1`

For the second partition:
- `/dev/sda2` (standard device node according to the SCSI device naming scheme)
- `/dev/disk/by-path/ccw-0.0.3c1b-fc-0x500507630300c562-lun-0x401040ea00000000-part2`
- `/dev/disk/by-uuid/b4a818c8-747c-40a2-bfa2-acaa3ef70ead`
- `/dev/disk/by-label/SWAP-sda2`
- `/dev/disk/by-id/scsi-36005076303ffc562000000000000010ea-part2`
- `/dev/disk/by-id/wwn-0x6005076303ffc562000000000000010ea-part2`

Device nodes by-uuid use a unique file-system identifier that does not relate to the partition number.
**Multipath**

Users of SCSI-over-Fibre Channel attached devices should always consider setting up and using redundant paths through their Fibre Channel storage area network.

Path redundancy improves the availability of the LUNs. In Linux, you can set up path redundancy with the device-mapper multipath tool. For information about multipath devices and multipath partitions, see the chapter about multipathing in *How to use FC-attached SCSI devices with Linux on z Systems*, SC33-8413.

**Partitioning a SCSI device**

You can partition SCSI devices that are attached through an FCP channel in the same way that you can partition SCSI attached devices on other platforms.

**About this task**

Use the `fdisk` command to partition a SCSI disk, not `fdasd`.

udev creates device nodes for partitions automatically. For the SCSI disk `/dev/sda`, the partition device nodes are called `/dev/sda1`, `/dev/sda2`, `/dev/sda3`, and so on.

**Example**

To partition a SCSI disk with a device node `/dev/sda` issue:

```
# fdisk /dev/sda
```

**zfcp HBA API (FC-HBA) support**

The zfcp host bus adapter API (HBA API) provides an interface for HBA management clients that run on Z.

As shown in Figure 38 on page 164 the zfcp HBA API support includes a user space library.
The SNIA (Storage Networking Industry Association) library can interface with the zFCP HBA API. The SNIA library is part of Red Hat Enterprise Linux 7.5. It is available as software package libhbaapi, which is automatically installed when installing the zfcp HBA API, see “Getting ready to run applications” on page 198.

The default method in Red Hat Enterprise Linux 7.5 is for applications to use the zFCP HBA API library indirectly through the SNIA HBA API. If you develop applications yourself, see “Developing applications” on page 196.

In a Linux on Z environment HBAs are usually virtualized and are shown as FCP devices. FCP devices are represented by CCW devices that are listed in /sys/bus/ccw/drivers/zfcp. Do not confuse FCP devices with SCSI devices. A SCSI device is a disk device that is identified by a LUN.

For information about setting up the HBA API support, see “zfcp HBA API support” on page 196.

**N_Port ID Virtualization for FCP channels**

Through N_Port ID Virtualization (NPIV), the sole port of an FCP channel appears as multiple, distinct ports with separate port identification.

NPIV support can be configured on the SE per CHPID and LPAR for an FCP channel. The zfcp device driver supports NPIV error messages and adapter attributes. See “Displaying FCP channel and device information” on page 168 for the Fibre Channel adapter attributes.

For more information, see the connectivity page at [www.ibm.com/systems/z/connectivity](http://www.ibm.com/systems/z/connectivity)
See also the chapter on NPIV in *How to use FC-attached SCSI devices with Linux on z Systems, SC33-8413.*

**Setting up the zfcp device driver**

Red Hat Enterprise Linux 7 loads the zfcp device driver for you when an FCP device becomes available. Use anaconda or dracut to configure the zfcp device driver. You might also need to install the zfcp HBA API library.

**Important:** Configuration changes can directly or indirectly affect information that is required to mount the root file system. Such changes require an update of the initial RAM disk, followed by a re-write of the boot record (see “Rebuilding the initial RAM disk image” on page 84).

You have the following options for configuring FCP LUNs to attach SCSI devices:

- During installation, use the anaconda GUI, the dracut boot parameter `rd.zfcp=`, or the kickstart parameter `zfcp`.
- On an installed system, use the dracut boot parameter `rd.zfcp=` only for SCSI disks that are required for the root file system. Use the configuration file `/etc/zfcp.conf` for all other SCSI devices, such as data volumes or tape libraries.

For details, see *Red Hat Enterprise Linux 7.5 Installation Guide*

You can always specify additional zfcp module parameters as explained in Chapter 3, “Kernel and module parameters,” on page 23.

### zfcp module parameter syntax

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Modprobe Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>allow_lun_scan</td>
<td>allow_lun_scan=1</td>
</tr>
<tr>
<td>datarouter</td>
<td>datarouter=1</td>
</tr>
<tr>
<td></td>
<td>datarouter=0</td>
</tr>
<tr>
<td>dbflevel</td>
<td>dbflevel=3</td>
</tr>
<tr>
<td></td>
<td>dbflevel=&lt;level&gt;</td>
</tr>
<tr>
<td>dbfsiz e</td>
<td>dbfsiz e=4</td>
</tr>
<tr>
<td></td>
<td>dbfsiz e=&lt;pages&gt;</td>
</tr>
<tr>
<td>dif</td>
<td>dif=0</td>
</tr>
<tr>
<td></td>
<td>dif=1</td>
</tr>
<tr>
<td>port_scan_ratelimit</td>
<td>port_scan_ratelimit=60000</td>
</tr>
<tr>
<td></td>
<td>port_scan_ratelimit=&lt;limit&gt;</td>
</tr>
<tr>
<td>port_scan_backoff</td>
<td>port_scan_backoff=500</td>
</tr>
<tr>
<td></td>
<td>port_scan_backoff=&lt;limit&gt;</td>
</tr>
<tr>
<td>no_auto_port_rescan</td>
<td>no_auto_port_rescan=0</td>
</tr>
<tr>
<td></td>
<td>no_auto_port_rescan=1</td>
</tr>
<tr>
<td>queue_depth</td>
<td>queue_depth=32</td>
</tr>
<tr>
<td></td>
<td>queue_depth=&lt;depth&gt;</td>
</tr>
</tbody>
</table>

where:

**allow_lun_scan=value**

disables the automatic LUN scan for FCP devices that run in NPIV mode if set to 0, n, or N. To enable the LUN scanning set the parameter to 1, y, or Y. When the LUN scan is disabled, all LUNs must be configured through the unit_add zfcp attribute in sysfs. LUN scan is enabled by default.
**datarouter=**  
enables (if set to 1, y, or Y) or disables (if set to 0, n, or N) support for the hardware data routing feature. The default is 1.

**Note:** The hardware data routing feature becomes active only for FCP devices that are based on adapter hardware with hardware data routing support.

**dbflevel=<level>**  
sets the initial log level of the debug feature. The value is an integer in the range 0 - 6, where greater numbers generate more detailed information. The default is 3.

**dbfsize=<pages>**  
specifies the number of pages to be used for the debug feature.

The debug feature is available for each FCP device and the following areas:

- **hba**  
  FCP device
- **san**  
  Storage Area Network
- **rec**  
  Error Recovery Process
- **scsi**  
  SCSI
- **pay**  
  Payloads for entries in the hba, san, rec, or scsi areas. The default is 8 pages.

The value given is used for all areas. The default is 4, that is, four pages are used for each area and FCP device. In the following example the dbfsize is increased to 6 pages:

```
dbfsize=6
```

This results in six pages being used for each area and FCP device. The payload is doubled to use 12 pages.

**dif=**  
turns end-to-end data consistency checking on if set to 1, y, or Y and off if set to 0, n, or N. The default is 0.

**no_auto_port_rescan=**  
turns the automatic port rescan feature off (if set to 1, y, or Y) or on (if set to 0, n, or N). The default is 0. Automatic rescan is always performed when setting an adapter online and when user-triggered writes to the sysfs attribute port_rescan occur.

**port_scan_ratelimit=<limit>**  
sets the minimum delay, in milliseconds, between automatic port scans of your Linux instance. The default value is 60000 milliseconds. To turn off the rate limit, specify 0. Use this parameter to avoid frequent scans, while you still ensure that a scan is conducted eventually.

**port_scan_backoff=<delay>**  
sets additional random delay, in milliseconds, in which the port scans of your Linux instance are spread. The default value is 500 milliseconds. To turn off the random delay, specify 0. In an installation with multiple Linux instances, use this attribute for every Linux instance to spread scans to avoid potential multiple simultaneous scans.

**queue_depth=<depth>**  
specifies the number of commands that can be issued simultaneously to a SCSI device. The default is 32. The value you set here is used as the default queue depth for new SCSI devices. You can set the queue depth for each SCSI device using the queue_depth sysfs attribute, see “Setting the queue depth” on page 188.

**device=<device_bus_id>, <wwpn>, <fcp_lun>**

**Attention:** The device= module parameter is reserved for internal use. Do not use.
<device_bus_id>
specifies the FCP device through which the SCSI device is attached.

<wwpn>
specifies the target port through which the SCSI device is attached.

<fcp_lun>
specifies the LUN of the SCSI device.

---

**Working with FCP devices**

Set an FCP device online before you attempt to perform any other tasks.

Working with FCP devices comprises the following tasks:

- “Setting an FCP device online or offline”
- “Displaying FCP channel and device information” on page 168
- “Recovering a failed FCP device” on page 172
- “Finding out whether NPIV is in use” on page 173
- “Logging I/O subchannel status information” on page 174

**Setting an FCP device online or offline**

By default, FCP devices are offline. Set an FCP device online before you perform any other tasks.

**About this task**

**Attention:** Use the procedure described here for dynamic testing of configuration settings. For persistent configuration in a production system, use one of the following options:

- The configuration file `/etc/zipl.conf` for FCP devices that are part of the root file system
- The configuration file `/etc/zfcp.conf` for data disks.

An example of how to define an FCP device persistently is in *Red Hat Enterprise Linux 7.5 Installation Guide*. For a general discussion of configuration files, see *Red Hat Enterprise Linux 7.5 System Administrator’s Guide*.

You have the following options for configuring FCP LUNs to attach SCSI devices:

- During installation, use the anaconda GUI, the dracut boot parameter `rd.zfcp=`, or the kickstart parameter `zfcp`.
- On an installed system, use the dracut boot parameter `rd.zfcp=` only for SCSI disks that are required for the root file system. Use the configuration file `/etc/zfcp.conf` for all other SCSI devices, such as data volumes or tape libraries.

For details, see *Red Hat Enterprise Linux 7.5 Installation Guide*

You can always specify additional zfcp module parameters as explained in Chapter 3, “Kernel and module parameters,” on page 23

**Important:** Configuration changes can directly or indirectly affect information that is required to mount the root file system. Such changes require an update of the initial RAM disk, followed by a re-write of the boot record (see “Rebuilding the initial RAM disk image” on page 84).
See “Working with newly available devices” on page 10 to avoid errors when you work with devices that have become available to a running Linux instance.

Setting an FCP device online registers it with the Linux SCSI stack. It also automatically runs the scan for ports in the SAN and waits for this port scan to complete. To check if setting the FCP device online was successful, you can use a script that first sets the FCP device online and after this operation completes checks if the WWPN of a remote storage port has appeared in sysfs.

When you set an FCP device offline, the port and LUN subdirectories are preserved. Setting an FCP device offline in sysfs interrupts the communication between Linux and the FCP channel. After a timeout has expired, the port and LUN attributes indicate that the ports and LUNs are no longer accessible. The transition of the FCP device to the offline state is synchronous, unless the device is disconnected.

For disconnected devices, writing 0 to the online sysfs attribute triggers an asynchronous deregistration process. When this process is completed, the device with its ports and LUNs is no longer represented in sysfs.

When the FCP device is set back online, the SCSI device names and minor numbers are freshly assigned. The mapping of devices to names and numbers might be different from what they were before the FCP device was set offline.

Procedure

There are two methods for setting an FCP device online or offline:

- Use the chccwdev command ("chccwdev - Set CCW device attributes” on page 490).

- Alternatively, you can write 1 to an FCP device’s online attribute to set it online, or 0 to set it offline.

Examples

- To set an FCP device with bus ID 0.0.3d0c online issue:

  ```
  # chccwdev -e 0.0.3d0c
  
  or
  # echo 1 > /sys/bus/ccw/drivers/zfcp/0.0.3d0c/online
  ```

- To set an FCP device with bus ID 0.0.3d0c offline issue:

  ```
  # chccwdev -d 0.0.3d0c
  
  or
  # echo 0 > /sys/bus/ccw/drivers/zfcp/0.0.3d0c/online
  ```

Displaying FCP channel and device information

For each online FCP device, there is a number of read-only attributes in sysfs that provide information about the corresponding FCP channel and FCP device.
Before you begin

The FCP device must be online for the FCP channel information to be valid.

About this task

The following tables summarize the relevant attributes.

Table 19. Attributes with Fibre Channel adapter hardware information

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>card_version</td>
<td>Version number that identifies a particular hardware feature.</td>
</tr>
<tr>
<td>hardware_version</td>
<td>Number that identifies a hardware version for a particular feature. The initial hardware version of a feature is zero. This version indicator is increased only for hardware modifications of the same feature. Appending hardware_version to card_version results in a hierarchical version indication for a physical adapter.</td>
</tr>
<tr>
<td>lic_version</td>
<td>Microcode level.</td>
</tr>
<tr>
<td>peer_wwnn</td>
<td>WWNN of peer for a point-to-point connection.</td>
</tr>
<tr>
<td>peer_wwpn</td>
<td>WWPN of peer for a point-to-point connection.</td>
</tr>
<tr>
<td>peer_d_id</td>
<td>Destination ID of the peer for a point-to-point connection.</td>
</tr>
</tbody>
</table>

Table 20. Attributes with FCP device information

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>in_recovery</td>
<td>Shows if the FCP device is in recovery (0 or 1).</td>
</tr>
</tbody>
</table>

For the attributes availability, cmb_enable, and cutype, see “Device directories” on page 8. The status attribute is reserved.

Table 21. Relevant transport class attributes, fc_host attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxframe_size</td>
<td>Maximum frame size of adapter.</td>
</tr>
<tr>
<td>node_name</td>
<td>Worldwide node name (WWNN) of adapter.</td>
</tr>
<tr>
<td>permanent_port_name</td>
<td>WWPN associated with the physical port of the FCP channel.</td>
</tr>
<tr>
<td>port_id</td>
<td>A unique ID (N_PORT_ID) assigned by the fabric. In an NPIV setup, each virtual port is assigned a different port_id.</td>
</tr>
<tr>
<td>port_name</td>
<td>WWPN associated with the FCP device. If N_PORT ID Virtualization is not available, the WWPN of the physical port (see permanent_port_name).</td>
</tr>
<tr>
<td>port_type</td>
<td>The port type indicates the topology of the port.</td>
</tr>
<tr>
<td>serial_number</td>
<td>The 32-byte serial number of the adapter hardware that provides the FCP channel.</td>
</tr>
<tr>
<td>speed</td>
<td>Speed of FC link.</td>
</tr>
<tr>
<td>supported_classes</td>
<td>Supported FC service class.</td>
</tr>
<tr>
<td>supported_speeds</td>
<td>Supported speeds.</td>
</tr>
<tr>
<td>tgid_bind_type</td>
<td>Target binding type.</td>
</tr>
</tbody>
</table>
Table 22. Relevant transport class attributes, fc_host statistics

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>reset_statistics</td>
<td>Writeable attribute to reset statistic counters.</td>
</tr>
<tr>
<td>seconds_since_last_reset</td>
<td>Seconds since last reset of statistic counters.</td>
</tr>
<tr>
<td>tx_frames</td>
<td>Transmitted FC frames.</td>
</tr>
<tr>
<td>tx_words</td>
<td>Transmitted FC words.</td>
</tr>
<tr>
<td>rx_frames</td>
<td>Received FC frames.</td>
</tr>
<tr>
<td>rx_words</td>
<td>Received FC words.</td>
</tr>
<tr>
<td>lip_count</td>
<td>Number of LIP sequences.</td>
</tr>
<tr>
<td>nos_count</td>
<td>Number of NOS sequences.</td>
</tr>
<tr>
<td>error_frames</td>
<td>Number of frames that are received in error.</td>
</tr>
<tr>
<td>dumped_frames</td>
<td>Number of frames that are lost because of lack of host resources.</td>
</tr>
<tr>
<td>link_failure_count</td>
<td>Link failure count.</td>
</tr>
<tr>
<td>loss_of_sync_count</td>
<td>Loss of synchronization count.</td>
</tr>
<tr>
<td>loss_of_signal_count</td>
<td>Loss of signal count.</td>
</tr>
<tr>
<td>prim_seq_protocol_err_count</td>
<td>Primitive sequence protocol error count.</td>
</tr>
<tr>
<td>invalid_tx_word_count</td>
<td>Invalid transmission word count.</td>
</tr>
<tr>
<td>invalid_crc_count</td>
<td>Invalid CRC count.</td>
</tr>
<tr>
<td>fcp_input_requests</td>
<td>Number of FCP operations with data input.</td>
</tr>
<tr>
<td>fcp_output_requests</td>
<td>Number of FCP operations with data output.</td>
</tr>
<tr>
<td>fcp_control_requests</td>
<td>Number of FCP operations without data movement.</td>
</tr>
<tr>
<td>fcp_input_megabytes</td>
<td>Megabytes of FCP data input.</td>
</tr>
<tr>
<td>fcp_output_megabytes</td>
<td>Megabytes of FCP data output.</td>
</tr>
</tbody>
</table>

Procedure

Use the `cat` command to read an attribute.

- Issue a command of this form to read an attribute:

```
# cat /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<attribute>
```

where:

- `<device_bus_id>` specifies an FCP device that corresponds to the FCP channel.
- `<attribute>` is one of the attributes in Table 19 on page 169 or Table 20 on page 169

- To read attributes of the associated Fibre Channel host use:

```
# cat /sys/class/fc_host/<host_name>/<attribute>
```

where:

- `<host_name>` is the ID of the Fibre Channel host.
- `<attribute>` is one of the attributes in Table 21 on page 169
To read statistics attributes of the FCP channel associated with this Fibre Channel host, use:

```
# cat /sys/class/fc_host/<host_name>/statistics/<attribute>
```

where:

- `<host_name>` is the ID of the Fibre Channel host.
- `<attribute>` is one of the attributes in Table 22 on page 170.

Examples

- In this example, information is displayed about an FCP channel that corresponds to an FCP device with bus ID 0.0.3d0c:

```
# cat /sys/bus/ccw/drivers/zfcp/0.0.3d0c/hardware_version
0x00000000
# cat /sys/bus/ccw/drivers/zfcp/0.0.3d0c/lic_version
0x00000715
```

- Alternatively you can use `lszfcp` (see “lszfcp - List zfcp devices” on page 595) to display attributes of an FCP channel:
Failed FCP devices are automatically recovered by the zfcp device driver. You can read the in_recovery attribute to check whether recovery is under way.

Before you begin

The FCP device must be online.
**Procedure**

Perform these steps to find out the recovery status of an FCP device and, if needed, start or restart recovery:

1. Issue a command of this form:
   ```
   # cat /sys/bus/ccw/drivers/zfcp/<device_bus_id>/in_recovery
   ```
   The value is 1 if recovery is under way and 0 otherwise. If the value is 0 for a non-operational FCP device, recovery might have failed. Alternatively, the device driver might have failed to detect that the FCP device is malfunctioning.

2. To find out whether recovery failed, read the failed attribute. Issue a command of this form:
   ```
   # cat /sys/bus/ccw/drivers/zfcp/<device_bus_id>/failed
   ```
   The value is 1 if recovery failed and 0 otherwise.

3. You can start or restart the recovery process for the FCP device by writing 0 to the failed attribute. Issue a command of this form:
   ```
   # echo 0 > /sys/bus/ccw/drivers/zfcp/<device_bus_id>/failed
   ```

**Example**

In the following example, an FCP device with a device bus-ID 0.0.3d0c is malfunctioning. The first command reveals that recovery is not already under way. The second command manually starts recovery for the FCP device:

```
# cat /sys/bus/ccw/drivers/zfcp/0.0.3d0c/in_recovery
0
# echo 0 > /sys/bus/ccw/drivers/zfcp/0.0.3d0c/failed
```

**Finding out whether NPIV is in use**

An FCP device runs in NPIV mode if the port_type attribute of the FCP device attribute contains the string "NPIV". Alternatively, if the applicable permanent_port_name and port_name are not the same and are not NULL.

**Procedure**

Read the port_type attribute of the FCP device. Issue a command of the following form:

```
# cat /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<host_no>/fc_host/<host_no>/port_type
```

For example:

```
# cat /sys/bus/ccw/drivers/zfcp/0.0.3d0c/host0/fc_host/host0/port_type
NPIV VPORT
```

Alternatively, compare the values of the permanent_port_name attribute and the port_name.
Tip: You can use `lszfcp` (see “lszfcp - List zfcp devices” on page 595) to list the FCP device attributes.

### Example

```bash
$ lszfcp -b 0.0.3d0c -a
0.0.3d0c host0
  Bus = "ccw"
  availability = "good"
  ...
Class = "fc_host"
  node_name = "0x5005076400c7ec87"
  permanent_port_name = "0xc05076ffd6801981"
  port_id = "0x671a29"
  port_name = "0xc05076ffd6801e10"
  port_state = "Online"
  port_type = "NPIV VPORT"
  ...
  symbolic_name = "IBM type serial PCHID: 0198 UlpId: 05600300 DEVNO: 0.0.3d0c NAME: hostname.domain"
  ...
```

The `port_type` attribute directly indicates that NPIV is used. The example also shows that `permanent_port_name` is different from `port_name` and neither is NULL. The example also shows the `symbolic_name` attribute that shows the symbolic port name that was registered on the FC name server.

#### Logging I/O subchannel status information

When severe errors occur for an FCP device, the FCP device driver triggers a set of log entries with I/O subchannel status information.

The log entries are available through the SE Console Actions Work Area with the View Console Logs function. In the list of logs, these entries have the prefix `1F00`. The content of the entries is intended for support specialists.

### Working with target ports

You can scan for ports, display port information, recover a port, or remove a port.

Working with target ports comprises the following tasks:

- “Scanning for ports” on page 174
- “Controlling automatic port scanning” on page 175
- “Displaying port information” on page 178
- “Recovering a failed port” on page 179
- “Removing ports” on page 180

#### Scanning for ports

Newly available target ports are discovered. However, you might want to trigger a port scan to re-create accidentally removed port information or to assure that all ports are present.

Before you begin

The FCP device must be online.
About this task

The zfcp device driver automatically adds port information to sysfs when:

- The FCP device is set online
- Target ports are added to the Fibre Channel fabric, unless the module parameter `no_auto_port_rescan` is set to 1. See "Setting up the zfcp device driver" on page 165.

Scanning for ports might take some time to complete. Commands that you issue against ports or LUNs while scanning is in progress are delayed and processed when port scanning is completed.

Use the `port_rescan` attribute if a remote storage port was accidentally deleted from the adapter configuration or if you are unsure whether all ports were added to sysfs.

Procedure

Issue a command of this form:

```
# echo 1 > /sys/bus/ccw/drivers/zfcp/<device_bus_id>/port_rescan
```

where `<device_bus_id>` specifies the FCP device through which the target ports are attached.

**Tip:** List the contents of `/sys/bus/ccw/drivers/zfcp/<device_bus_id>` to find out which ports are currently configured for the FCP device.

Example

In this example, a port with WWPN 0x500507630303c562 is already configured for an FCP device with bus ID 0.0.3d0c. An additional target port with WWPN 0x500507630300c562 is automatically configured by triggering a port scan.

```
# ls /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x
0x500507630303c562
# echo 1 > /sys/bus/ccw/drivers/zfcp/0.0.3d0c/port_rescan
# ls /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x
0x500507630303c562
0x500507630300c562
```

Controlling automatic port scanning

Automatic port scanning includes two zfcp parameters that improve the behaviour of Linux instances in SANs. These zfcp parameters are set to default values that work well for most installations.

If needed, you can fine-tune the frequency and timing of automatic port scans with the zfcp parameters `port_scan_backoff` and `port_scan_ratelimit`.

You can enable automatic port scanning with the zfcp parameter `no_auto_port_rescan=0`. This value is the default.

About this task

In a large installation, where many Linux instances receive the same notifications of SAN changes, multiple instances might trigger scans simultaneously and too
frequently. See [Figure 39]

![Figure 39. Numerous port scans in a Linux installation](image)

These scans might put unnecessary load on the name server function of fabric switches and potentially result in late or inconclusive results.

You can avoid excessive scanning, yet still ensure that a port scan is eventually conducted. You can control port scanning with the zfcp parameters:

- **port_scan_ratelimit**
  - sets the minimum delay, in milliseconds, between automatic port scans of your Linux instance. The default value is 60000 milliseconds. To turn off the rate limit, specify 0.

- **port_scan_backoff**
  - sets an additional random delay, in milliseconds, in which the port scans of your Linux instance are spread. In an installation with multiple Linux instances, use this zfcp parameter for every Linux instance to spread scans to avoid potential multiple simultaneous scans. The default value is 500 milliseconds. To turn off the random delay, specify 0.

Use module parameters (see “Setting up the zfcp device driver” on page 165). On a running Linux system, you can also query or set these values by using the sysfs attributes with the same names.

Using **port_scan_ratelimit** reduces the number of scans, as shown in [Figure 40]

![Figure 40. Port scan behavior with scan rate limit.](image)

However, if the rate limit is set to the same value, the scans can still occur almost simultaneously, as for FCP device A and B in Linux 1.

Using **port_scan_backoff** and **port_scan_ratelimit** together delays port scans even further and avoids simultaneous scans, as shown in [Figure 41 on page 177] In
the figure, FCP devices A and B in Linux 1 have the same rate limit and the same backoff values. The random element in the backoff value causes the scans to occur at slightly different times.

![Diagram showing port scan behavior with backoff and scan rate limit.](image)

**Figure 41. Port scan behavior with backoff and scan rate limit.**

**Procedure**

Use `port_scan_backoff` and `port_scan_ratelimit` together or separately to tune the behavior of port scanning:

- To avoid too frequent scanning, set a minimum wait time between two consecutive scans for the same Linux instance. Use the `port_scan_ratelimit` sysfs attribute. By default, `port_scan_ratelimit` is turned on and has a value of 60000 milliseconds. For example, to specify an attribute value of 12 seconds, issue:

  ```sh
  # echo 12000 > /sys/module/zfcp/parameters/port_scan_ratelimit
  ```

- To further spread scans over a certain time and thus avoid multiple simultaneous scans, set the `port_scan_backoff` sysfs attribute. By default, `port_scan_backoff` is turned on and has a value of 500 milliseconds. For example, to query the setting, issue a command of this form:

  ```sh
  # cat /sys/module/zfcp/parameters/port_scan_backoff
  500
  ```

  To set the attribute to 1 second, issue:

  ```sh
  # echo 1000 > /sys/module/zfcp/parameters/port_scan_backoff
  ```

**Results**

The automatic port scans are delayed by the values specified. If a SAN notification is received during the rate limit time, a port scan is conducted immediately after the delay time passed.

Setting the attributes in sysfs is a useful method on a running system where you want to make dynamic changes. If you want to make the changes persistent across IPLs, use the kernel or module parameter.

Depending on the port event, one or more of the three zfcp parameters are evaluated to schedule a port scan. For example, port scans that are triggered manually through sysfs are not delayed. Table 23 on page 178 shows which events evaluate which zfcp parameters.
### Displaying port information

For each target port, there is a number of read-only sysfs attributes with port information.

#### About this task

[Table 24](#) and [Table 25](#) summarize the relevant attributes.

#### Table 24. zfcp-specific attributes with port information within the FCP device sysfs tree

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>access_denied</td>
<td>This attribute is obsolete. The value is always 0.</td>
</tr>
<tr>
<td>in_recovery</td>
<td>Shows if port is in recovery (0 or 1).</td>
</tr>
</tbody>
</table>

#### Table 25. Transport class attributes with port information

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>node_name</td>
<td>WWNN of the remote port (target port).</td>
</tr>
<tr>
<td>port_name</td>
<td>WWPN of remote port.</td>
</tr>
<tr>
<td>port_id</td>
<td>Destination ID of remote port.</td>
</tr>
<tr>
<td>port_state</td>
<td>State of remote port.</td>
</tr>
<tr>
<td>roles</td>
<td>Role of remote port (usually FCP target).</td>
</tr>
<tr>
<td>scsi_target_id</td>
<td>Linux SCSI ID of remote port.</td>
</tr>
<tr>
<td>supported_classes</td>
<td>Supported classes of service.</td>
</tr>
</tbody>
</table>

#### Procedure

Use the `cat` command to read an attribute.

- Issue a command of this form to read a zfcp-specific attribute:

  ```
  # cat /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>/<attribute>
  ```

  where:
<device_bus_id>
specifies the FCP device.

<wwpn>
is the WWPN of the target port.

<attribute>
is one of the attributes in Table 24 on page 178.

- To read transport class attributes of the associated target port, use a command of this form:

```
# cat /sys/class/fc_remote_ports/<rport_name>/<attribute>
```

where:

<rport_name>
is the name of the remote port.

<attribute>
is one of the attributes in Table 25 on page 178.

Tip: With the HBA API package installed, you can also use the `zfcp ping` and `zfcp show` commands to find out more about your ports. See “Tools for investigating your SAN configuration” on page 198.

Examples

- In this example, information is displayed for a target port 0x500507630300c562 that is attached through an FCP device with bus ID 0.0.3d0c:

```
# cat /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562/in_recovery
```

- To display transport class attributes of a target port you can use `lszfcp`:

```
# lszfcp -p 0x500507630300c562 -a
0.0.3d0c/0x500507630300c562 rport-0:0-0
Class = "fc_remote_ports"
  dev_loss_tmo   = "2147483647"
  fast_io_fail_tmo = "5"
  maxframe_size  = "2048 bytes"
  node_name      = "0x500507630300c562"
  port_id        = "0x652113"
  port_name      = "0x500507630300c562"
  port_state     = "Online"
  roles          = "FCP Target"
  scsi_target_id = "0"
  supported_classes = "Class 2, Class 3"
```

Recovering a failed port

Failed target ports are automatically recovered by the zfcp device driver. You can read the `in_recovery` attribute to check whether recovery is under way.

Before you begin

The FCP device must be online.

Procedure

Perform these steps to find out the recovery status of a port and, if needed, start or restart recovery:

1. Issue a command of this form:
where:
<device_bus_id>
    specifies the FCP device.
<wwpn>
    is the WWPN of the target port.
The value is 1 if recovery is under way and 0 otherwise. If the value is 0 for a non-operational port, recovery might have failed or the device driver might have failed to detect that the port is malfunctioning.

2. To find out whether recovery failed, read the failed attribute. Issue a command of this form:

```
# cat /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>/failed
```

The value is 1 if recovery has failed and 0 otherwise.

3. You can start or restart the recovery process for the port by writing 0 to the failed attribute. Issue a command of this form:

```
# echo 0 > /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>/failed
```

Example

In the following example, a port with WWPN 0x500507630300c562 that is attached through an FCP device with bus ID 0.0.3d0c is malfunctioning. The first command reveals that recovery is not already under way. The second command manually starts recovery for the port:

```
# cat /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562/in_recovery
0
# echo 0 > /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562/failed
```

Removing ports

Removing unused ports can save FCP channel resources. Additionally setting the no_auto_port_rescan attribute avoids unnecessary attempts to recover unused remote ports.

Before you begin

The FCP device must be online.

About this task

List the contents of /sys/bus/ccw/drivers/zfcp/<device_bus_id> to find out which ports are currently configured for the FCP device.

You cannot remove a port while SCSI devices are configured for it (see "Configuring SCSI devices" on page 182) or if the port is in use, for example, by error recovery.
Note: The next port scan will attach all available ports, including any previously removed ports. To prevent removed ports from being reattached automatically, use zoning or the no_auto_port_rescan module parameter, see “Setting up the zfcp device driver” on page 165.

Procedure

To remove a port from an FCP device, write the WWPN of the port to the port_remove attribute of the FCP device.

Issue a command of this form:

```
# echo <wwpn> > /sys/bus/ccw/drivers/zfcp/<device_bus_id>/port_remove
```

where:

- `<device_bus_id>` specifies the FCP device.
- `<wwpn>` is the WWPN of the port to be removed.

Example

In this example, two ports with WWPN 0x500507630303c562 and 0x500507630300c562 are configured for an FCP device with bus ID 0.0.3d0c. The port with WWPN 0x500507630303c562 is then removed.

```
# ls /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x*
0x500507630303c562
0x500507630300c562
# echo 0x500507630303c562 > /sys/bus/ccw/drivers/zfcp/0.0.3d0c/port_remove
# ls /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x*
0x500507630300c562
```

Working with SCSI devices

In an NPIV setup with auto lun scan, the SCSI devices are configured automatically. Otherwise, you must configure FCP LUNs to obtain SCSI devices. In both cases, you can configure SCSI devices, display information, and remove SCSI devices.

Working with SCSI devices comprises the following tasks:

- “Configuring SCSI devices” on page 182
- “Mapping the representations of a SCSI device in sysfs” on page 184
- “Displaying information about SCSI devices” on page 185
- “Setting the queue depth” on page 188
- “Recovering failed SCSI devices” on page 189
- “Updating the information about SCSI devices” on page 190
- “Setting the SCSI command timeout” on page 190
- “Controlling the SCSI device state” on page 191
- “Removing SCSI devices” on page 192
Configuring SCSI devices

FCP devices that use NPIV mode detect the LUNs automatically and no configuring is necessary. If needed, configure the LUN manually.

For each FCP device that uses NPIV mode and if you did not disable automatic LUN scanning (see "Setting up the zfcp device driver" on page 165), the LUNs are configured for you. In this case, no FCP LUN entries are created under /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>.

To find out whether an FCP device is using NPIV mode, check the port_type attribute, for example:

```
# cat /sys/bus/ccw/drivers/zfcp/0.0.1901/host0/fc_host/host0/port_type
NPIV VPORT
```

To find out whether automatic LUN scanning is enabled, check the current setting of the module parameter zfcp.allow_lun_scan. The example below shows automatic LUN scanning as turned on.

```
# cat /sys/module/zfcp/parameters/allow_lun_scan
Y
```

Important: Configuration changes can directly or indirectly affect information that is required to mount the root file system. Such changes require an update of the initial RAM disk, followed by a re-write of the boot record (see "Rebuilding the initial RAM disk image" on page 84).

Automatically attached SCSI devices

FCP devices that use NPIV mode detect the LUNs automatically and no configuring is necessary.

In this case, no FCP LUN entries are created under /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>.

What to do next

To check whether a SCSI device is registered, check for a directory with the name of the LUN in /sys/bus/scsi/devices. If there is no SCSI device for this LUN, the LUN is not valid in the storage system, or the FCP device is offline in Linux.

Manually configured FCP LUNs and their SCSI devices

For FCP devices that do not use NPIV mode, or if automatic LUN scanning is disabled, FCP LUNs must be configured manually to obtain SCSI devices.

Before you begin

Attention: Use this procedure only to dynamically test configuration settings.

You have the following options for configuring FCP LUNs to attach SCSI devices:

- During installation, use the anaconda GUI, the dracut boot parameter rd.zfcp=, or the kickstart parameter zfcp.
On an installed system, use the `dracut` boot parameter `rd.zfcp=` only for SCSI disks that are required for the root file system. Use the configuration file `/etc/zfcp.conf` for all other SCSI devices, such as data volumes or tape libraries.

For details, see *Red Hat Enterprise Linux 7.5 Installation Guide*

You can always specify additional zfcp module parameters as explained in Chapter 3, “Kernel and module parameters,” on page 23

**Procedure**

If your FCP device does not use NPIV mode, or if you have disabled automatic LUN scanning, proceed as follows:

To configure a SCSI device for a target port, write the device's LUN to the port's `unit_add` attribute. Issue a command of this form:

```
# echo <fcp_lun> > /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>/unit_add
```

where:

- `<fcp_lun>` is the LUN of the SCSI device to be configured. The LUN is a 16 digit hexadecimal value padded with zeroes, for example `0x4010403300000000`.
- `<device_bus_id>` specifies the FCP device.
- `<wwpn>` is the WWPN of the target port.

This command starts a process with multiple steps:

1. It creates a directory in `/sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>` with the LUN as the directory name. The directory is part of the list of all LUNs to configure. Without NPIV or with auto LUN scanning disabled, zfcp registers only FCP LUNs contained in this list with the Linux SCSI stack in the next step.
2. It initiates the registration of the SCSI device with the Linux SCSI stack. The FCP device must be online for this step.
3. It waits until the Linux SCSI stack registration completes successfully or returns an error. It then returns control to the shell. A successful registration creates a `sysfs` entry in the SCSI branch (see “Mapping the representations of a SCSI device in `sysfs`” on page 184).

**Example**

In this example, a target port with WWPN `0x500507630300c562` is attached through an FCP device with bus ID `0.0.3d0c`. A SCSI device with LUN `0x4010403200000000` is already configured for the port. An additional SCSI device with LUN `0x4010403300000000` is added to the port.

```
# ls /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562/0x
0x4010403200000000
# echo 0x4010403300000000 > /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562/unit_add
# ls /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562/0x
0x4010403200000000
0x4010403300000000
```
What to do next

To check whether a SCSI device is registered for the configured LUN, check for a directory with the name of the LUN in /sys/bus/scsi/devices. If there is no SCSI device for this LUN, the LUN is not valid in the storage system, or the FCP device is offline in Linux.

To see which LUNs are currently configured for the port, list the contents of /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>.

Mapping the representations of a SCSI device in sysfs

Each SCSI device that is configured is represented by multiple directories in sysfs, in particular, within the SCSI branch. Only manually configured LUNs are also represented within the zfcp branch.

About this task

The directory in the sysfs SCSI branch has the following form:

/sys/bus/scsi/devices/<scsi_host_no>:0:<scsi_id>:<scsi_lun>

where:

- `<scsi_host_no>` is the SCSI host number that corresponds to the FCP device.
- `<scsi_id>` is the SCSI ID of the target port.
- `<scsi_lun>` is the LUN of the SCSI device.

The values for `<scsi_id>` and `<scsi_lun>` depend on the storage device. Often, they are single-digit numbers but for some storage devices they have numerous digits.

For manually configured FCP LUNs, see "Manually configured FCP LUNs and their SCSI devices" on page 182 for details about the directory in the zfcp branch.

Figure 42 shows how the directory name is composed in the sysfs SCSI branch. The sysfs zfcp branch only exists for manually configured FCP LUNs. For manually configured FCP LUNs, the directory name is composed of attributes of consecutive directories and you can find the name of the directory in the sysfs SCSI branch by reading the corresponding attributes in the zfcp branch.

![Figure 42. SCSI devices in sysfs](image)

The hba_id, wwpn, and fcp_lun attributes of the SCSI device in the SCSI branch match the names of the `<device_bus_id>`, `<wwpn>`, and `<fcp_lun>` directories for the same SCSI device in the zfcp branch.
**Procedure**

Use `lszfcp` (see “lszfcp - List zfcp devices” on page 595) to map the two representations of a SCSI device.

**Example**

This example shows how you can use `lszfcp` to display the name of the SCSI device that corresponds to a zfcp unit, for example:

```bash
# lszfcp -1 0x4010403200000000
0.0.3d0c/0x500507630300c562/0x4010403200000000 0:0:0:0
```

In the example, the output informs you that the unit with the LUN 0x4010403200000000, which is configured on a port with the WWPN 0x500507630300c562 for an FCP device with bus ID 0.0.3d0c, maps to SCSI device "0:0:0:0".

To confirm that the SCSI device belongs to the zfcp unit:

```bash
# cat /sys/bus/scsi/devices/0:0:0:0/hba_id
0.0.3d0c
# cat /sys/bus/scsi/devices/0:0:0:0/wwpn
0x500507630300c562
# cat /sys/bus/scsi/devices/0:0:0:0/fcp_lun
0x4010403200000000
```

**Displaying information about SCSI devices**

For each SCSI device, there is a number of read-only attributes in sysfs that provide information for the device.

**About this task**

Table 26 summarizes the read-only attributes for manually configured FCP LUNs, including those attributes that indicate whether the device access is restricted by access control software on the FCP channel. These attributes can be found in the zfcp branch of sysfs. The path has the form:

```
/sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>/<fcp_lun>/<attribute>
```

**Table 26. Attributes of manually configured FCP LUNs with device access information**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>access_denied</td>
<td>Flag that indicates whether access to the device is restricted by the FCP channel. The value is 1 if access is denied and 0 if access is permitted. If access is denied to your Linux instance, confirm that your SCSI devices are configured as intended. Also, be sure that you really want to share a SCSI device. For shared access to a SCSI device, preferably use NPIV (see “N_Port ID Virtualization for FCP channels” on page 164). You might also use different FCP channels or target ports.</td>
</tr>
<tr>
<td>access_shared</td>
<td>This attribute is obsolete. The value is always 0.</td>
</tr>
<tr>
<td>access_readonly</td>
<td>This attribute is obsolete. The value is always 0.</td>
</tr>
<tr>
<td>in_recovery</td>
<td>Shows if unit is in recovery (0 or 1)</td>
</tr>
</tbody>
</table>
Table 27 lists further read-only attributes with information about the SCSI device. These attributes can be found in the SCSI branch of sysfs. The path has the form:

/sy/s/class/scsi_device/<device_name>/device/<attribute>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>device_blocked</td>
<td>Flag that indicates whether the device is in blocked state (0 or 1).</td>
</tr>
<tr>
<td>iocounterbits</td>
<td>The number of bits used for I/O counters.</td>
</tr>
<tr>
<td>iodone_cnt</td>
<td>The number of completed or rejected SCSI commands.</td>
</tr>
<tr>
<td>ioerr_cnt</td>
<td>The number of SCSI commands that completed with an error.</td>
</tr>
<tr>
<td>iorequest_cnt</td>
<td>The number of issued SCSI commands.</td>
</tr>
<tr>
<td>queue_type</td>
<td>The type of queue for the SCSI device. The value can be one of the following:</td>
</tr>
<tr>
<td></td>
<td>• none</td>
</tr>
<tr>
<td></td>
<td>• simple</td>
</tr>
<tr>
<td></td>
<td>• ordered</td>
</tr>
<tr>
<td>model</td>
<td>The model of the SCSI device, received from inquiry data.</td>
</tr>
<tr>
<td>rev</td>
<td>The revision of the SCSI device, received from inquiry data.</td>
</tr>
<tr>
<td>scsi_level</td>
<td>The SCSI revision level, received from inquiry data.</td>
</tr>
<tr>
<td>type</td>
<td>The type of the SCSI device, received from inquiry data.</td>
</tr>
<tr>
<td>vendor</td>
<td>The vendor of the SCSI device, received from inquiry data.</td>
</tr>
<tr>
<td>fcp_lun</td>
<td>The LUN of the SCSI device in 64-bit format.</td>
</tr>
<tr>
<td>hba_id</td>
<td>The bus ID of the SCSI device.</td>
</tr>
<tr>
<td>wwpn</td>
<td>The WWPN of the remote port.</td>
</tr>
<tr>
<td>zfcp_access_denied</td>
<td>Flag that indicates whether access to the device is restricted by the FCP channel.</td>
</tr>
<tr>
<td></td>
<td>The value is 1 if access is denied and 0 if access is permitted.</td>
</tr>
<tr>
<td>zfcp_in_recovery</td>
<td>Shows if unit is in recovery (0 or 1).</td>
</tr>
</tbody>
</table>

**Procedure**

Issue a command of this form to read an attribute of a manually configured FCP LUN:

```
# cat /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>/<fcp_lun>/<attribute>
```

where:

- `<device_bus_id>` specifies the FCP device.
- `<wwpn>` is the WWPN of the target port.
- `<fcp_lun>` is the FCP LUN of the SCSI device.
is one of the attributes in Table 26 on page 185.

Use the `lszfcp` command (see "lszfcp - List zfcp devices" on page 595) to display information about the associated SCSI device. Alternatively, you can use sysfs to read the information. To read attributes of the associated SCSI device, use a command of this form:

```bash
# cat /sys/class/scsi_device/<device_name>/device/<attribute>
```

where:

- `<device_name>` is the name of the associated SCSI device.
- `<attribute>` is one of the attributes in Table 27 on page 186.

**Tip:** For SCSI-attached tape devices, you can display a summary of this information by using the `lstape` command (see "lstape - List tape devices" on page 587).

**Examples**

- In this example, information is displayed for a manually configured FCP LUN with LUN 0x4010403200000000 that is accessed through a target port with WWPN 0x500507630300c562 and is attached through an FCP device 0.0.3d0c. For the device, access is permitted.

  ```bash
  # cat /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562/0x4010403200000000/access_denied
  0
  ```

  For the device to be accessible, the `access_denied` attribute of the target port, 0x500507630300c562, must also be 0 (see "Displaying port information" on page 178).

- You can use `lszfcp` to display attributes of a SCSI device. The example shows the attributes listed in Table 27 on page 186 as well as other relevant attributes:
Setting the queue depth

The Linux SCSI code automatically adjusts the queue depth as necessary. Changing the queue depth is usually a storage server requirement.

Before you begin

Check the documentation of the storage server used or contact your storage server support group to establish if there is a need to change this setting.

About this task

The value of the queue_depth kernel parameter (see "Setting up the zfcp device driver" on page 165) is used as the default queue depth of new SCSI devices. You can query the queue depth by issuing a command of this form:

```
# cat /sys/bus/scsi/devices/<SCSI device>/queue_depth
```

Example:

```
# cat /sys/bus/scsi/devices/0:0:19:1086537744/queue_depth
16
```

You can change the queue depth of each SCSI device by writing to the queue_depth attribute, for example:

```
# echo 8 > /sys/bus/scsi/devices/0:0:19:1086537744/queue_depth
# cat /sys/bus/scsi/devices/0:0:19:1086537744/queue_depth
8
```
This is useful on a running system where you want to make dynamic changes. If you want to make the changes persistent across IPLs, you can:

- Use the module parameter queue_depth described in “Setting up the zfcp device driver” on page 165.
- Write a udev rule to change the setting for each new SCSI device.

Linux forwards SCSI commands to the storage server until the number of pending commands exceeds the queue depth. If the server lacks the resources to process a SCSI command, Linux queues the command for a later retry and decreases the queue depth counter. Linux then waits for a defined ramp-up period. If no indications of resource problems occur within this period, Linux increases the queue depth counter until reaching the previously set maximum value. To query the current value for the queue ramp-up period in milliseconds:

```
# cat /sys/bus/scsi/devices/0:0:13:1086537744/queue_ramp_up_period
120000
```

To set a new value for the queue ramp-up period in milliseconds:

```
# echo 1000 > /sys/bus/scsi/devices/0:0:13:1086537744/queue_ramp_up_period
```

### Recovering failed SCSI devices

Failed SCSI devices are automatically recovered by the zfcp device driver. You can read the zfcp_in_recovery attribute to check whether recovery is under way.

#### Before you begin

The FCP device must be online.

#### Procedure

Perform the following steps to check the recovery status of a failed SCSI device:

1. Check the value of the zfcp_in_recovery attribute. Issue the `lszfcp` command:

   ```
   # lszfcp -l <LUN> -a
   ```

   where `<LUN>` is the LUN of the associated SCSI device.

   Alternatively, you can issue a command of this form:

   ```
   # cat /sys/class/scsi_device/<device_name>/device/zfcp_in_recovery
   ```

   The value is 1 if recovery is under way and 0 otherwise. If the value is 0 for a non-operational SCSI device, recovery might have failed. Alternatively, the device driver might have failed to detect that the SCSI device is malfunctioning.

2. To find out whether recovery failed, read the zfcp_failed attribute. Either use the `lszfcp` command again, or issue a command of this form:

   ```
   # cat /sys/class/scsi_device/<device_name>/device/zfcp_failed
   ```

   The value is 1 if recovery has failed, and 0 otherwise.

3. You can start or restart the recovery process for the SCSI device by writing \(0\) to the zfcp_failed attribute. Issue a command of this form:
Example

In the following example, SCSI device 0:0:0 is malfunctioning. The first command reveals that recovery is not already under way. The second command manually starts recovery for the SCSI device:

```bash
# cat /sys/class/scsi_device/0:0:0:0/device/zfcp_in_recovery
0
# echo 0 > /sys/class/scsi_device/0:0:0:0/device/zfcp_failed
```

What to do next

If you manually configured an FCP LUN (see “Manually configured FCP LUNs and their SCSI devices” on page 182), but did not get a corresponding SCSI device, you can also use the corresponding FCP LUN sysfs attributes, in_recovery and failed, to check on recovery. See Table 26 on page 185.

Updating the information about SCSI devices

Use the rescan attribute of the SCSI device to detect changes to a storage device on the storage server that are made after the device was discovered.

Before you begin

The FCP device must be online.

About this task

The initial information about the available SCSI devices is discovered automatically when LUNs first become available.

Procedure

To update the information about a SCSI device issue a command of this form:

```bash
# echo <string> > /sys/bus/scsi/devices/<scsi_host_no>:0:<scsi_id>:<scsi_lun>/rescan
```

where <string> is any alphanumeric string and the other variables have the same meaning as in “Mapping the representations of a SCSI device in sysfs” on page 184.

Example

In the following example, the information about a SCSI device 1:0:18:1086537744 is updated:

```bash
# echo 1 > /sys/bus/scsi/devices/1:0:18:1086537744/rescan
```

Setting the SCSI command timeout

You can change the timeout if the default is not suitable for your storage system.
Before you begin

The FCP device must be online.

About this task

There is a timeout for SCSI commands. If the timeout expires before a SCSI command completes, error recovery starts. The default timeout is 30 seconds.

To find out the current timeout, read the timeout attribute of the SCSI device:

```
# cat /sys/bus/scsi/devices/<scsi_host_no>:0:<scsi_id>:<scsi_lun>/timeout
```

where the variables have the same meaning as in “Mapping the representations of a SCSI device in sysfs” on page 184.

The attribute value specifies the timeout in seconds.

Procedure

To set a different timeout, enter a command of this form:

```
# echo <timeout> > /sys/bus/scsi/devices/<scsi_host_no>:0:<scsi_id>:<scsi_lun>/timeout
```

where `<timeout>` is the new timeout in seconds.

Example

In the following example, the timeout of a SCSI device 1:0:18:1086537744 is first read and then set to 45 seconds:

```
# cat /sys/bus/scsi/devices/1:0:18:1086537744/timeout
30
# echo 45 > /sys/bus/scsi/devices/1:0:18:1086537744/timeout
```

Controlling the SCSI device state

You can use the state attribute of the SCSI device to set a SCSI device back online if it was set offline by error recovery.

Before you begin

The FCP device must be online.

About this task

If the connection to a storage system is working but the storage system has a problem, the error recovery might set the SCSI device offline. This condition is indicated by a message like “Device offline - not ready after error recovery”.

To find out the current state of the device, read the state attribute:

```
# cat /sys/bus/scsi/devices/<scsi_host_no>:0:<scsi_id>:<scsi_lun>/state
```
where the variables have the same meaning as in the "Mapping the representations of a SCSI device in sysfs" on page 184. The state can be:

- **running**
  The SCSI device can be used for running regular I/O requests.

- **cancel**
  The data structure for the device is being removed.

- **deleted**
  Follows the cancel state when the data structure for the device is being removed.

- **quiesce**
  No I/O requests are sent to the device, only special requests for managing the device. This state is used when the system is suspended.

- **offline**
  Error recovery for the SCSI device has failed.

- **blocked**
  Error recovery is in progress and the device cannot be used until the recovery process is completed.

**Procedure**

To set an offline device online again, write `running` to the state attribute.

Issue a command of this form:

```
# echo running > /sys/bus/scsi/devices/<scsi_host_no>:0:<scsi_id>:<scsi_lun>/state
```

**Example**

In the following example, SCSI device `1:0:18:1086537744` is offline and is then set online again:

```
# cat /sys/bus/scsi/devices/1:0:18:1086537744/state
offline
# echo running > /sys/bus/scsi/devices/1:0:18:1086537744/state
```

**Removing SCSI devices**

How to remove a SCSI device depends on whether your environment is set up to use NPIV.

**Important:** Configuration changes can directly or indirectly affect information that is required to mount the root file system. Such changes require an update of the initial RAM disk, followed by a re-write of the boot record (see "Rebuilding the initial RAM disk image" on page 84).

**Removing automatically attached SCSI devices**

When running with NPIV and automatic LUN scan, you can temporarily delete a SCSI device by writing `1` to the `delete` attribute of the directory that represents the device in the `sysfs` SCSI branch.

**About this task**

See "Mapping the representations of a SCSI device in sysfs" on page 184 about how to find this directory.

**Note:** These steps delete the SCSI device only temporarily, until the next automatic or user triggered Linux SCSI target scan. The scan automatically adds the SCSI
devices again, unless the LUNs were deconfigured on the storage target. To permanently delete SCSI devices, you must disable automatic LUN scanning and configure all LUNs manually, see “Manually configured FCP LUNs and their SCSI devices” on page 182.

**Procedure**

Issue a command of this form:

```
# echo 1 > /sys/bus/scsi/devices/<device>/delete
```

**Example**

In this example, an SCSI device with LUN 0x4010403700000000 is to be removed. Before the device is deleted, the corresponding device in the sysfs SCSI branch is found with an `lszfcp` command.

```
# lszfcp -1 0x4010403700000000
0.0.3d0f/0x500507630300c567/0x4010403700000000 0:0:3:1
# echo 1 > /sys/bus/scsi/devices/0:0:3:1/delete
```

**Removing manually configured FCP LUNs and their SCSI device**

Manually remove a SCSI device if your environment is not set up to use NPIV or if you disabled automatic LUN scan.

For details about disabling automatic LUN scan, see “Setting up the zfcp device driver” on page 165.

**Before you begin**

**Attention:** Use this procedure only to dynamically test configuration settings.

To configure a persistent setting in a production system, remove the dracut boot parameter `rd.zfcp=` only for SCSI disks that are no longer required for the root file system. Use the configuration file `/etc/zfcp.conf` for all other SCSI devices, such as data volumes or tape libraries.

**Procedure**

Follow these steps to remove a manually configured FCP LUN and its SCSI device:

1. **Optional:** To manually unregister the SCSI device, write 1 to the `delete` attribute of the directory that represents the device in the sysfs SCSI branch. See “Mapping the representations of a SCSI device in sysfs” on page 184 for information about how to find this directory. Issue a command of this form:

```
# echo 1 > /sys/bus/scsi/devices/<device>/delete
```

2. Remove the SCSI device from the target port by writing the LUN of the device to the `unit_remove` attribute of the port. Issue a command of this form:

```
# echo <fcp_lun> > /sys/bus/ccw/drivers/zfcp/<device_bus_id>/<wwpn>/unit_remove
```

where the variables have the same meaning as in “Configuring SCSI devices” on page 182. Removing a LUN with `unit_remove` automatically unregisters the SCSI device first.
Example

The following example removes a SCSI device with LUN 0x4010403200000000, accessed through a target port with WWPN 0x500507630300c562 and is attached through an FCP device with bus ID 0.0.3d0c. The corresponding directory in the sysfs SCSI branch is assumed to be /sys/bus/scsi/devices/0:0:1:1.

1. Optionally, unregister the device:

   ```
   # echo 1 > /sys/bus/scsi/devices/0:0:1:1/delete
   ```

2. Remove the device (if not done in previous step) and the LUN:

   ```
   # echo 0x4010403200000000 > /sys/bus/ccw/drivers/zfcp/0.0.3d0c/0x500507630300c562/unit_remove
   ```

Confirming end-to-end data consistency checking

There are different types of end-to-end data consistency checking, with dependencies on hardware and software.

About this task

End-to-end data consistency checking is based on a data integrity field (DIF) that is added to transferred data blocks. DIF data is used to confirm that a data block originates from the expected source and was not modified during the transfer between the storage system and the FCP device. The SCSI Block Commands (T10 SBC) standard defines several types of DIF. Linux data integrity extension (DIX) builds on DIF to extend consistency checking, for example, to the operating system, middleware, or an application.

You enable the zfcp device driver for end-to-end data consistency checking with the `zfcp.dif=kernel` or `dif=module` parameter (see “Setting up the zfcp device” on page 165). With end-to-end data consistency checking for SCSI disks enabled, Linux automatically discovers which FCP devices and which SCSI disks support end-to-end data consistency checking. No further setup is required.

Note: SCSI devices for which end-to-end data consistency checking is enabled must be accessed with direct I/O. Direct I/O requires direct access through the block device or through a file system that fully supports end-to-end data consistency checking. For example, XFS provides this support. Expect error messages about invalid checksums when using other access methods.

The zfcp device driver supports the following modes:

- The FCP device calculates and checks a DIF checksum (DIF type 1)
- The Linux block integrity layer calculates and checks a TCP/IP checksum, which the FCP device then translates to a DIF checksum (DIX type 1 with DIF type 1)

For SCSI devices for which end-to-end data consistency checking is used, there is a sysfs directory

/sys/block/sd<x>/integrity

In the path, `sd<x>` is the standard name of the SCSI device.

End-to-end data consistency checking is used only if all of the following components support it:
SCSI disk
Check your storage server documentation about T10 DIF support and any restrictions.

IBM Z hardware
IBM Z FCP adapter hardware supports end-to-end data consistency checking as of FICON Express8.

Hypervisor
For Linux on z/VM, you require a z/VM version with guest support for end-to-end data consistency checking.

FCP device
Check your FCP adapter hardware documentation about the support and any restrictions. For example, end-to-end data consistency checking might be supported only for disks with 512-byte block size.

Read the prot_capabilities sysfs attribute of the SCSI host associated with an FCP device to find out about its end-to-end data consistency checking support. The following values are possible:

0  The FCP device does not support end-to-end data consistency checking.
1  The FCP device supports DIF type 1.
16 The FCP device supports DIX type 1.
17 The FCP device supports DIX type 1 with DIF type 1.

Procedure
Issue a command of this form:

```
# cat /sys/bus/ccw/devices/<device_bus_id>/host<n>/scsi_host/host<n>/prot_capabilities
```

where <device_bus_id> identifies the FCP device and <n> is an integer that identifies the corresponding SCSI host.

Example

```
# cat /sys/bus/ccw/devices/0.0.1940/host0/scsi_host/host0/prot_capabilities
17
```

Scenario for finding available LUNs
There are several steps from setting an FCP device online to listing the available LUNs.

Procedure
1. Check for available FCP devices of type 1732/03:

```
# lscss -t 1732/03
Device Subchan. DevType CU Type Use PIM PAM POM CHPIDs
-----------------------------------------------
0.0.3c02 0.0.0015 1732/03 1731/03 80 80 ff 36000000 00000000
```

Another possible type would be, for example, 1732/04.

2. Set the FCP device online:
A port scan is performed automatically when the FCP device is set online.

3. Optional: Confirm that the FCP device is available and online:

```bash
# lszfcp -b 0.0.3c02 -a 0.0.3c02 host0
Bus = "ccw"
   availability = "good"
   failed = "0"
   in_recovery = "0"
   online = "1"
```

4. Optional: List the available ports:

```bash
# lszfcp -P 0.0.3c02/0x50050763030bc562 rport-0:0-0
0.0.3c02/0x50050763030bc562 rport-0:0-1
0.0.3c02/0x50050763040727b rport-0:0-10
0.0.3c02/0x500507630e060521 rport-0:0-11
...```

5. Scan for available LUNs on FCP device 0.0.3c02, port 0x50050763030bc562:

```bash
# lsluns -c 0.0.3c02 -p 0x50050763030bc562
Scanning for LUNs on adapter 0.0.3c02 at port 0x50050763030bc562:
0x4010400000000000
0x4010400000000000
0x4010400000000000
0x4010400000000000
0x4010400000000000
...```

---

**zfcp HBA API support**

You require different libraries for developing and running HBA management client applications. To develop applications, you need the development version of the SNIA HBA API library. To run applications, you need the zFCP HBA API library.

**Developing applications**

To develop applications, you must install the development version of the SNIA HBA API provided by the `libhbaapi-devel` RPM, link your application against the library, and load the library.

**Procedure**

1. Install the development RPM for the SNIA HBA API. Use, for example, yum:

   ```bash
   # yum install libhbaapi-devel
   ```

   The development RPM `libhbaapi-devel` provides the necessary header files and .so symbolic links needed to program against the SNIA HBA API.
2. Add the command line option \texttt{-lhBAAPI} during the linker step of the build process to link your application against the SNIA HBA API library.

3. In the application, issue the \texttt{HBA\_LoadLibrary()} call as the first call to load the library. The vendor-specific library \texttt{libzfcpshaapi}, in turn, supplies the function \texttt{HBA\_RegisterLibrary} that returns all function pointers to the common library and thus makes them available to the application.

### Functions provided

The zfcp HBA API implements Fibre Channel - HBA API (FC-HBA) functions as defined in the FC-HBA specification.

You can find the FC-HBA specification at \url{www.t11.org}. The following functions are available:

- \texttt{HBA\_GetVersion()}
- \texttt{HBA\_LoadLibrary()}
- \texttt{HBA\_FreeLibrary()}
- \texttt{HBA\_RegisterLibrary()}
- \texttt{HBA\_RegisterLibraryV2()}
- \texttt{HBA\_GetNumberOfAdapters()}
- \texttt{HBA\_GetAdapterName()}
- \texttt{HBA\_OpenAdapter()}
- \texttt{HBA\_CloseAdapter()}
- \texttt{HBA\_RefreshInformation()}
- \texttt{HBA\_RefreshAdapterConfiguration()}
- \texttt{HBA\_GetAdapterAttributes()}
- \texttt{HBA\_GetAdapterPortAttributes()}
- \texttt{HBA\_GetDiscoveredPortAttributes()}
- \texttt{HBA\_GetFcpTargetMapping()}
- \texttt{HBA\_GetFcpTargetMappingV2()}
- \texttt{HBA\_SendScsiInquiry()}
- \texttt{HBA\_SendReadCapacity()}
- \texttt{HBA\_SendReportLUNs()}
- \texttt{HBA\_SendReportLUNsV2()}
- \texttt{HBA\_SendCTPassThru()}
- \texttt{HBA\_SendCTPassThruV2()}
- \texttt{HBA\_SetRNIDMgmtInfo()}
- \texttt{HBA\_GetRNIDMgmtInfo()}
- \texttt{HBA\_SendRNID()}
- \texttt{HBA\_SendRNIDV()}
- \texttt{HBA\_SendRPL()}
- \texttt{HBA\_SendRPS()}
- \texttt{HBA\_SendSRL()}
- \texttt{HBA\_SendLIRR()}
- \texttt{HBA\_GetEventBuffer()}
- \texttt{HBA\_RegisterForAdapterAddEvents()}
- \texttt{HBA\_RegisterForAdapterEvents()}
- \texttt{HBA\_RegisterForAdapterPortEvents()}
- \texttt{HBA\_RegisterForAdapterPortStatEvents()}
- \texttt{HBA\_RegisterForTargetEvents()}
- \texttt{HBA\_RegisterForLinkEvents()}
- \texttt{HBA\_RemoveCallback()}

All other FC-HBA functions return status code \texttt{HBA\_STATUS\_ERROR\_NOT\_SUPPORTED} where possible.
Note: ZFCP HBA API for Linux 3.10 can access only FCP devices, ports, and units that are configured in the operating system.

Getting ready to run applications

To run an application, you must install the zFCP HBA API library that is provided by the libzfcphbaapi RPM. You can set environment variables to log any errors in the library, and use tools to investigate the SAN configuration.

Before you begin

To use the HBA API support, you need the following packages:

- The zFCP HBA API library RPM, libzfcphbaapi.
- The SNIA HBA API library RPM, libhbaapi.

The application must be developed to use the SNIA HBA API library, see “Developing applications” on page 196.

Procedure

Follow these steps to access the library from a client application:

1. Install the libzfcphbaapi RPM with yum. Yum automatically installs all dependent packages such as libhbaapi. For example:

```
# yum install libzfcphbaapi
```

2. Ensure that the /etc/hba.conf file exists and contains a line of the form:

```
<library name> <library pathname>
```

For example:

```
libzfcphbaapi /usr/lib64/libzfcphbaapi-2.1.so
```

The SNIA library requires a configuration file called /etc/hba.conf that contains the path to the vendor-specific library libzfcphbaapi.so.

3. Optional: Set the environment variables for logging errors. The zfcp HBA API support uses the following environment variables to log errors in the zfcp HBA API library:

- **LIB_ZFCP_HBAAPI_LOG_LEVEL**
  - specifies the log level. If not set or set to zero, there is no logging (default). If set to an integer value greater than 1, logging is enabled.

- **LIB_ZFCP_HBAAPI_LOG_FILE**
  - specifies a file for the logging output. If not specified, stderr is used.

What to do next

You can use the zfcp_ping and zfcp_show commands to investigate your SAN configuration.

Tools for investigating your SAN configuration

As of version 2.1, the HBA API package includes tools that can help you to investigate your SAN configuration and to solve configuration problems.

- **zfcp_ping**
  - to probe a port in the SAN.
zfcp_show

to retrieve information about the SAN topology and details about the SAN components.

See *How to use FC-attached SCSI devices with Linux on z Systems, SC33-8413* for details.
Chapter 12. Storage-class memory device driver

Storage-class memory (SCM) is a class of data storage devices that combines properties of both storage and memory.

SCM can be implemented as Flash Express or as Virtual Flash Memory.

What you should know about storage-class memory

Storage-class memory (SCM) is accessed, in chunks called increments, through extended asynchronous data mover (EADM) subchannels.

The LPAR on which your Linux instance runs must be configured to provide SCM.

- At least one EADM subchannel must be available to the LPAR. Because SCM supports multiple concurrent I/O requests, it is advantageous to configure multiple EADM subchannels. A typical number of EADM subchannels is 64.
- One or more SCM increments must be added to the I/O configuration of the LPAR.

In Linux, each increment is represented as a block device. You can use the block device with standard Linux tools as you would use any other block device. Commonly used tools that work with block devices include: fdisk, mkfs, and mount.

Storage-class memory is useful for workloads with large write operations, that is, with a block size of 256 KB or more of data. Write operations with a block size of less than 256 KB of data might not perform optimally. Read operations can be of any size.

Storage-class memory device nodes

Applications access storage-class memory devices by device nodes. Red Hat Enterprise Linux creates a device node for each storage increment.

The device driver uses a device name of the form /dev/scm<x> for an entire block device. In the name, <x> is one or two lowercase letters.

You can partition a block device into up to seven partitions. If you use partitions, the device driver numbers them from 1 - 7. The partitions then have device nodes of the form /dev/scm<x><n>, where <n> is a number in the range 1 - 7, for example, /dev/scma1.

The following example shows two block devices, scma and scmb, where scma has one partition, scma1.

```
# lsblk
NAME   MAJ:MIN   RM  SIZE RO   MOUNTPOINT
scma   252:0     0  16G  0
~scma1  252:1     0  16G  0
scmb   252:8     0  16G  0
```
Setting up the storage-class memory device driver

Configure the storage-class memory device driver by using the module parameters.

### Storage-class memory module parameter syntax

```
modprobe scm_block
  nr_requests=64  write_cluster_size=64
  nr_requests=<num>  write_cluster_size=<num>

nr_request_per_io=8
  nr_request_per_io=<num>
```

where

- **nr_requests** specifies the number of parallel I/O requests. Set this number to the number of EADM subchannels. The default is 64.

- **write_cluster_size** specifies the number of pages that are used by the read-modify-write algorithm (available if CONFIG_SCM_BLOCK_CLUSTER_WRITE=y). The default is 64, resulting in all write requests smaller than 256 KiB being translated to 256 KiB writes. 1 KiB is 1024 bytes. Valid values are 2, 4, 8, 16, 32, 64, and 128. Do not change this value unless asked to do so by your service organization.

- **nr_request_per_io** submits more concurrent I/O requests than the current limit, which is based on the number of available EADM subchannels (64). Valid values are 1 - 64. Increasing the requests increases the number of I/O requests per second, especially for requests with a small block size. The default number of requests is 8. Depending on the workload, this setting might improve the throughput of the scm_block driver.

### Working with storage-class memory increments

You can list storage-class memory increments and EADM devices.

- "Displaying EADM subchannels"
- "Listing storage-class memory increments" on page 203
- "Combining SCM devices with LVM" on page 203

#### Displaying EADM subchannels

Use the **lscss** command to list EADM subchannels.

#### About this task

The extended asynchronous data mover (EADM) subchannels are used to transfer data to and from the storage-class memory. At least one EADM subchannel must be available to the LPAR.
Procedure

To list EADM subchannels, issue:

```
# lscss --eadm
Device Subchan.
--------- -------
n/a 0.0.ff00
n/a 0.0.ff01
n/a 0.0.ff02
n/a 0.0.ff03
n/a 0.0.ff04
n/a 0.0.ff05
n/a 0.0.ff06
n/a 0.0.ff07
```

For more information about the `lscss` command, see “lscss - List subchannels” on page 567.

Listing storage-class memory increments

Use the `lsscm` command to see the status and attributes of storage-class memory increments.

About this task

Each storage-class memory increment can be accessed as a block device through a device node `/dev/scm<x>`. Optionally, you can partition a storage-class memory increment in up to seven partitions.

You can also use the `lsblk` command to list all block devices.

Procedure

To list all storage-class memory increments, their status, and attributes, issue:

```
# lsscm
SCM Increment Size Name Rank D_state O_state Pers ResID
--------- ------- ------- ----- ------ ------ ------ ------
0000000000000000 16384MB scma 1 2 1 2 1
0000000000000000 16384MB scmb 1 2 1 2 1
```

See “lsscm - List storage-class memory increments” on page 584 for details about the `lsscm` command.

Combining SCM devices with LVM

You can use LVM to combine multiple SCM block devices into an arbitrary sized LVM device.

Example

Configure SCM as any other block devices in LVM. If your version of LVM does not accept SCM devices as valid LVM device types and issues an error message, add the SCM devices to the LVM configuration file `/etc/lvm/lvm.conf`. Add the following line to the section labeled “devices”:

```python
types = [ "scm", 8 ]
```
Chapter 13. Channel-attached tape device driver

The tape device driver supports channel-attached tape devices on Red Hat Enterprise Linux 7.5 for z Systems.

SCSI tape devices that are attached through an FCP channel are handled by the zfcp device driver (see Chapter 11, “SCSI-over-Fibre Channel device driver,” on page 159).

Features

The tape device driver supports a range of channel-attached tape devices and functions of these devices.

- The tape device driver supports channel-attached tape drives that are compatible with IBM 3480, 3490, 3590, and 3592 magnetic tape subsystems. Various models of these device types are handled (for example, the 3490/10).
- 3592 devices that emulate 3590 devices are recognized and treated as 3590 devices.
- Logical character devices for non-rewinding and rewinding modes of operation (see “Tape device modes and logical devices”).
- Control operations through mt (see “Using the mt command” on page 207).
- Message display support (see “tape390_display - Display messages on tape devices and load tapes” on page 623).
- Encryption support (see “tape390_crypt - Manage tape encryption” on page 619).
- Up to 128 physical tape devices.

What you should know about channel-attached tape devices

A naming scheme helps you to keep track of your tape devices, their modes of operation, and the corresponding device nodes.

Tape device modes and logical devices

The tape device driver supports up to 128 physical tape devices. Each physical tape device can be used as a character device in non-rewinding or in rewinding mode.

In non-rewinding mode, the tape remains at the current position when the device is closed. In rewinding mode, the tape is rewound when the device is closed. The tape device driver treats each mode as a separate logical device.

Both modes provide sequential (traditional) tape access without any caching done in the kernel.

You can use a channel-attached tape device in the same way as any other Linux tape device. You can write to it and read from it using standard Linux facilities such as GNU tar. You can perform control operations (such as rewinding the tape or skipping a file) with the standard tool mt.
Tape naming scheme

The tape device driver assigns minor numbers along with an index number when a physical tape device comes online.

The naming scheme for tape devices is summarized in Table 28:

Table 28. Tape device names and minor numbers

<table>
<thead>
<tr>
<th>Device</th>
<th>Names</th>
<th>Minor numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-rewinding character devices</td>
<td>ntibm&lt;n&gt;</td>
<td>2×&lt;n&gt;</td>
</tr>
<tr>
<td>Rewinding character devices</td>
<td>rtibm&lt;n&gt;</td>
<td>2×&lt;n&gt;+1</td>
</tr>
</tbody>
</table>

where <n> is the index number that is assigned by the device driver. The index starts from 0 for the first physical tape device, 1 for the second, and so on. The name space is restricted to 128 physical tape devices, so the maximum index number is 127 for the 128th physical tape device.

The index number and corresponding minor numbers and device names are not permanently associated with a specific physical tape device. When a tape device goes offline, it surrenders its index number. The device driver assigns the lowest free index number when a physical tape device comes online. An index number with its corresponding device names and minor numbers can be reassigned to different physical tape devices as devices go offline and come online.

Tip: Use the `lstape` command (see “lstape - List tape devices” on page 587) to determine the current mapping of index numbers to physical tape devices.

When the tape device driver is loaded, it dynamically allocates a major number to channel-attached character tape devices. A different major number might be used when the device driver is reloaded, for example when Linux is rebooted.

For online tape devices directories provide information about the major/minor assignments. The directories have the form:

- `/sys/class/tape390/ntibm<n>`
- `/sys/class/tape390/rtibm<n>`

Each of these directories has a `dev` attribute. The value of the `dev` attribute has the form `<major>:<minor>`, where `<major>` is the major number of the device and `<minor>` is the minor number specific to the logical device.

Example

In this example, four physical tape devices are present, with three of them online. The TapeNo column shows the index number and the BusID indicates the associated physical tape device. In the example, no index number is allocated to the tape device in the first row. The device is offline and, currently, no names and minor numbers are assigned to it.

```
# lstape --ccw-only
TapeNo BusID CuType/Model DevType/DevMod BlkSize State Op MedState
0 0.0.01a1 3490/10 3490/40 auto UNUSED --- UNLOADED
1 0.0.01a0 3480/01 3480/04 auto UNUSED --- UNLOADED
2 0.0.0172 3590/50 3590/11 auto IN_USE --- LOADED
N/A 0.0.01ac 3490/10 3490/40 N/A OFFLINE --- N/A
```

Table 29 on page 207 summarizes the resulting names and minor numbers.
Table 29. Example names and minor numbers

<table>
<thead>
<tr>
<th>Bus ID</th>
<th>Index (TapeNo)</th>
<th>Device</th>
<th>Device name</th>
<th>Minor number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.01a1</td>
<td>0</td>
<td>non-rewind</td>
<td>ntibm0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rewind</td>
<td>rtibm0</td>
<td>1</td>
</tr>
<tr>
<td>0.0.01a0</td>
<td>1</td>
<td>non-rewind</td>
<td>ntibm1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rewind</td>
<td>rtibm1</td>
<td>3</td>
</tr>
<tr>
<td>0.0.0172</td>
<td>2</td>
<td>non-rewind</td>
<td>ntibm2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rewind</td>
<td>rtibm2</td>
<td>5</td>
</tr>
<tr>
<td>0.0.01ac</td>
<td>not assigned</td>
<td>n/a</td>
<td>n/a</td>
<td>not assigned</td>
</tr>
</tbody>
</table>

For the online character devices, the major/minor assignments can be read from their respective representations in /sys/class:

```bash
# cat /sys/class/tape390/ntibm0/dev
254:0
# cat /sys/class/tape390/rtibm0/dev
254:1
# cat /sys/class/tape390/ntibm1/dev
254:2
# cat /sys/class/tape390/rtibm1/dev
254:3
# cat /sys/class/tape390/ntibm2/dev
254:4
# cat /sys/class/tape390/rtibm2/dev
254:5
```

In the example, the major number that is used for character devices is 254. The minor numbers are as expected for the respective device names.

**Tape device nodes**

Applications access tape devices by device nodes. Red Hat Enterprise Linux 7.5 uses udev to create two device nodes for each tape device.

The device nodes have the form /dev/<name>, where <name> is the device name according to "Tape naming scheme" on page 206.

For example, if you have two tape devices, udev creates the device nodes that are shown in Table 30.

**Table 30. Tape device nodes**

<table>
<thead>
<tr>
<th>Node for</th>
<th>non-rewind device</th>
<th>rewind device</th>
</tr>
</thead>
<tbody>
<tr>
<td>First tape device</td>
<td>/dev/ntibm0</td>
<td>/dev/rtibm0</td>
</tr>
<tr>
<td>Second tape device</td>
<td>/dev/ntibm1</td>
<td>/dev/rtibm1</td>
</tr>
</tbody>
</table>

**Using the mt command**

There are differences between the MTIO interface for channel-attached tapes and other tape drives. Correspondingly, some operations of the mt command are different for channel-attached tapes.

The mt command handles basic tape control in Linux. See the man page for general information about mt.
Basic Linux tape control is handled by the **mt** utility. See the man page for general information about **mt**.

**setdensity**

has no effect because the recording density is automatically detected on channel-attached tape hardware.

**drvbuffer**

has no effect because channel-attached tape hardware automatically switches to unbuffered mode if buffering is unavailable.

**lock / unlock**

have no effect because channel-attached tape hardware does not support media locking.

**setpartition / mkpartition**

have no effect because channel-attached tape hardware does not support partitioning.

**status**

returns a structure that, aside from the block number, contains mostly SCSI-related data that does not apply to the tape device driver.

**load**

does not automatically load a tape but waits for a tape to be loaded manually.

**offline or rewoffl or eject**

all include expelling the currently loaded tape. Depending on the stacker mode, it might attempt to load the next tape (see “Loading and unloading tapes” on page 212 for details).

---

**Loading the tape device driver**

You must load the appropriate tape device driver module before you can work with tape devices.

Use the **modprobe** command to ensure that any other required modules are loaded in the correct order.

**Tape module syntax**

```
modprobe tape_34xx
modprobe tape_3590
```

See the **modprobe** man page for details about **modprobe**.

To load the tape device driver module automatically at boot time, see the section on persistent module loading in *Red Hat Enterprise Linux 7.5 System Administrator’s Guide*.

**Working with tape devices**

Typical tasks for working with tape devices include displaying tape information, controlling compression, and loading and unloading tapes.
For information about working with the channel measurement facility, see Chapter 45, “Channel measurement facility,” on page 457.

For information about displaying messages on a tape device’s display unit, see “tape390_display - Display messages on tape devices and load tapes” on page 623.

See “Working with newly available devices” on page 10 to avoid errors when working with devices that have become available to a running Linux instance.

- “Setting a tape device online or offline”
- “Displaying tape information” on page 210
- “Enabling compression” on page 212
- “Loading and unloading tapes” on page 212

## Setting a tape device online or offline

Set a tape device online or offline with the chccwdev command or through the online sysfs attribute of the device.

### About this task

Setting a physical tape device online makes both corresponding logical devices accessible:

- The non-rewind character device
- The rewind character device

At any time, the device can be online to a single Linux instance only. You must set the tape device offline to make it accessible to other Linux instances in a shared environment.

### Procedure

Use the chccwdev command (see “chccwdev - Set CCW device attributes” on page 490) to set a tape online or offline.

Alternatively, you can write 1 to the device's online attribute to set it online or 0 to set it offline.

### Results

When a physical tape device is set online, the device driver assigns an index number to it. This index number is used in the standard device nodes (see “Tape device nodes” on page 207) to identify the corresponding logical devices. The index number is in the range 0 - 127. A maximum of 128 physical tape devices can be online concurrently.

If you are using the standard device nodes, you must find out which index number the tape device driver has assigned to your tape device. This index number, and consequently the associated standard device node, can change after a tape device was set offline and back online.

If you need to know the index number, issue a command of this form:

```bash
# lstape --ccw-only <device_bus_id>
```

where `<device_bus_id>` is the device bus-ID that corresponds to the physical tape device. The index number is the value in the TapeNo column of the command.
output. For more information about the `lstape` command, see “List tape devices” on page 587.

Examples

- To set a physical tape device with device bus-ID 0.0.015f online, issue:

  ```
  # chccwdev -e 0.0.015f
  ```

  or

  ```
  # echo 1 > /sys/bus/ccw/devices/0.0.015f/online
  ```

To find the index number that the tape device driver assigned to the device, issue:

```
# lstape 0.0.015f --ccw-only
```

<table>
<thead>
<tr>
<th>TapeNo</th>
<th>BusID</th>
<th>CuType/Model</th>
<th>DevType/Model</th>
<th>BlkSize</th>
<th>State</th>
<th>Op</th>
<th>MedState</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0.015f</td>
<td>3480/01</td>
<td>3480/04</td>
<td>auto</td>
<td>UNUSED</td>
<td>---</td>
<td>LOADED</td>
<td></td>
</tr>
</tbody>
</table>

In the example, the assigned index number is 2. The standard device nodes for working with the device until it is set offline are then:

- `/dev/ntibm2` for the non-rewinding device
- `/dev/rtibm2` for the rewinding device

- To set a physical tape device with device bus-ID 0.0.015f offline, issue:

  ```
  # chccwdev -d 0.0.015f
  ```

  or

  ```
  # echo 0 > /sys/bus/ccw/devices/0.0.015f/online
  ```

Displaying tape information

Use the `lstape` command to display summary information about your tape devices, or read tape information from `sysfs`.

Alternatively, you can read tape information from `sysfs`. Each physical tape device is represented in a `sysfs` directory of the form

```
/sys/bus/ccw/devices/<device_bus_id>
```

where `<device_bus_id>` is the device bus-ID that corresponds to the physical tape device. This directory contains a number of attributes with information about the physical device. The attributes: `blocksize`, `state`, `operation`, and `medium_state`, might not show the current values if the device is offline.

**Table 31. Tape device attributes**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>online</td>
<td>1 if the device is online or 0 if it is offline (see <a href="#">“Setting a tape device online or offline”</a> on page 209)</td>
</tr>
<tr>
<td>cmb_enable</td>
<td>1 if channel measurement block is enabled for the physical device or 0 if it is not enabled (see <a href="#">Chapter 45, “Channel measurement facility,”</a> on page 457)</td>
</tr>
</tbody>
</table>
Table 31. Tape device attributes (continued)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>cutype</td>
<td>Type and model of the control unit</td>
</tr>
<tr>
<td>devtype</td>
<td>Type and model of the physical tape device</td>
</tr>
<tr>
<td>blocksize</td>
<td>Currently used block size in bytes or 0 for auto</td>
</tr>
<tr>
<td>state</td>
<td>State of the physical tape device, either of:</td>
</tr>
<tr>
<td></td>
<td>UNUSED</td>
</tr>
<tr>
<td></td>
<td>Device is not in use and is available to any operating system image in a shared environment</td>
</tr>
<tr>
<td></td>
<td>IN_USE</td>
</tr>
<tr>
<td></td>
<td>Device is being used as a character device by a process on this Linux image</td>
</tr>
<tr>
<td></td>
<td>OFFLINE</td>
</tr>
<tr>
<td></td>
<td>The device is offline.</td>
</tr>
<tr>
<td></td>
<td>NOT_OP</td>
</tr>
<tr>
<td></td>
<td>Device is not operational</td>
</tr>
<tr>
<td>operation</td>
<td>The current tape operation, for example:</td>
</tr>
<tr>
<td></td>
<td>--- No operation</td>
</tr>
<tr>
<td></td>
<td>WRI Write operation</td>
</tr>
<tr>
<td></td>
<td>RFO Read operation</td>
</tr>
<tr>
<td></td>
<td>MSN Medium sense</td>
</tr>
<tr>
<td></td>
<td>Several other operation codes exist, for example, for rewind and seek.</td>
</tr>
<tr>
<td>medium_state</td>
<td>The current state of the tape cartridge:</td>
</tr>
<tr>
<td>1</td>
<td>Cartridge is loaded into the tape device</td>
</tr>
<tr>
<td>2</td>
<td>No cartridge is loaded</td>
</tr>
<tr>
<td>0</td>
<td>The tape device driver does not have information about the current cartridge state</td>
</tr>
</tbody>
</table>

Procedure

Issue a command of this form to read an attribute:

```
# cat /sys/bus/ccw/devices/<device_bus_id>/<attribute>
```

where `<attribute>` is one of the attributes of Table 31 on page 210.

Example

The following `1stape` command displays information about a tape device with bus ID 0.0.015f:

```
# 1stape 0.0.015f --ccw-only
```

<table>
<thead>
<tr>
<th>TapeNo</th>
<th>BusID</th>
<th>CuType/Model</th>
<th>DevType/Model</th>
<th>BlkSize</th>
<th>State</th>
<th>Op</th>
<th>MedState</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0.015f</td>
<td>3480/01</td>
<td>3480/04</td>
<td>auto</td>
<td>UNUSED</td>
<td>---</td>
<td>LOADED</td>
</tr>
</tbody>
</table>

This sequence of commands reads the same information from sysfs:
Enabling compression

Control Improved Data Recording Capability (IDRC) compression with the `mt` command provided by the RPM mt-st.

About this task

Compression is off after the tape device driver is loaded.

Procedure

To enable compression, issue:

```
# mt -f <node> compression
```

or

```
# mt -f <node> compression 1
```

where `<node>` is the device node for a character device, for example, `/dev/ntibm0`. To disable compression, issue:

```
# mt -f <tape> compression 0
```

Any other numeric value has no effect, and any other argument disables compression.

Example

To turn on compression for a tape device with a device node `/dev/ntibm0` issue:

```
# mt -f /dev/ntibm0 compression 1
```

Loading and unloading tapes

Unload tapes with the `mt` command. How to load tapes depends on the stacker mode of your tape hardware.
Procedure

Unload tapes by issuing a command of this form:

```
# mt -f <node> unload
```

where `<node>` is one of the character device nodes.

Whether or not you can load tapes from your Linux instance depends on the stacker mode of your tape hardware. There are three possible modes:

**manual**

Tapes must always be loaded manually by an operator. You can use the `tape390_display` command (see “`tape390_display - Display messages on tape devices and load tapes`” on page 623) to display a short message on the tape device's display unit when a new tape is required.

**automatic**

If there is another tape present in the stacker, the tape device automatically loads a new tape when the current tape is expelled. You can load a new tape from Linux by expelling the current tape with the `mt` command.

**system**

The tape device loads a tape when instructed from the operating system. From Linux, you can load a tape with the `tape390_display` command (see “`tape390_display - Display messages on tape devices and load tapes`” on page 623). You cannot use the `mt` command to load a tape.

Example

To expel a tape from a tape device that can be accessed through a device node `/dev/ntibm0`, issue:

```
# mt -f /dev/ntibm0 unload
```

Assuming that the stacker mode of the tape device is “system” and that a tape is present in the stacker, you can load a new tape by issuing:

```
# tape390_display -l "NEW TAPE" /dev/ntibm0
```

“NEW TAPE” is a message that is displayed on the tape devices display unit until the tape device receives the next tape movement command.
Chapter 14. XPRAM device driver

With the XPRAM block device driver Red Hat Enterprise Linux 7.5 for z Systems can access expanded storage. XPRAM can be used as a basis for fast swap devices or for fast file systems.

Expanded storage can be swapped in or out of the main storage in 4 KB blocks. All XPRAM devices provide a block size of 4096 bytes.

XPRAM features

The XPRAM device driver automatically detects expanded storage and supports expanded storage partitions.

- If expanded storage is not available, XPRAM fails gracefully with a log message that reports the absence of expanded storage.
- The expanded storage can be divided into up to 32 partitions.

What you should know about XPRAM

There is a device node for each XPRAM partition. Expanded storage persists across reboots and, with suitable boot parameters, the stored data can be accessed by the rebooted Linux instance.

XPRAM partitions and device nodes

The XPRAM device driver uses major number 35. The standard device names are of the form slram<n>, where <n> is the corresponding minor number.

You can use the entire available expanded storage as a single XPRAM device or divide it into up to 32 partitions. Each partition is treated as a separate XPRAM device.

If the entire expanded storage is used a single device, the device name is slram0. For partitioned expanded storage, the <n> in the device name denotes the (n+1)th partition. For example, the first partition is called slram0, the second slram1, and the 32nd partition is called slram31.

Table 32. XPRAM device names, minor numbers, and partitions

<table>
<thead>
<tr>
<th>Minor</th>
<th>Name</th>
<th>To access</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>slram0</td>
<td>the first partition or the entire expanded storage if there are no partitions</td>
</tr>
<tr>
<td>1</td>
<td>slram1</td>
<td>the second partition</td>
</tr>
<tr>
<td>2</td>
<td>slram2</td>
<td>the third partition</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>&lt;n&gt;</td>
<td>slram&lt;n&gt;</td>
<td>the (&lt;n&gt;+1)th partition</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>31</td>
<td>slram31</td>
<td>the 32nd partition</td>
</tr>
</tbody>
</table>

The device nodes that you need to access these partitions are created by udev when you load the XPRAM device driver module. The nodes are of the form

© Copyright IBM Corp. 2000, 2018
/dev/s1ram<n>, where <n> is the index number of the partition. In addition, to the device nodes udev creates a symbolic link of the form /dev/xpram<n> that points to the respective device node.

**XPRAM use for diagnosis**

Expanded storage persists across reboots, which makes it suitable for storing diagnostic information.

Issuing an IPL command to reboot Linux does not reset expanded storage. Expanded storage is persistent across IPLs and can be used, for example, to store diagnostic information. The expanded storage is reset when the z/VM guest virtual machine is logged off or when the LPAR is activated.

**Reusing XPRAM partitions**

You might be able to reuse existing file systems or swap devices on an XPRAM device or partition after reloading the XPRAM device driver (for example, after rebooting Linux).

For file systems or swap devices to be reusable, the XPRAM kernel or module parameters for the new device or partition must match the parameters of the previous use of XPRAM.

If you change the XPRAM parameters, you must create a new file system or a new swap device for each changed partition. A device or partition is considered changed if its size has changed. All partitions that follow a changed partition are also considered changed even if their sizes are unchanged.

**Setting up the XPRAM device driver**

You can optionally partition the available expanded storage by using the devs and sizes module parameters when you load the xpram module.

```
modprobe xpram
  dev=<number_of_partitions>,
  sizes=<partition_size>
```

where:

* `<number_of_partitions>`
  is an integer in the range 1 - 32 that defines how many partitions the expanded storage is split into.

* `<partition_size>`
  specifies the size of a partition. The i-th value defines the size of the i-th partition.

Each size is a non-negative integer that defines the size of the partition in KB or a blank. Only decimal values are allowed and no magnitudes are accepted.
You can specify up to `<number_of_partitions>` values. If you specify fewer values than `<number_of_partitions>`, the missing values are interpreted as blanks. Blanks are treated like zeros.

Any partition that is defined with a non-zero size is allocated the amount of memory that is specified by its size parameter.

Any remaining memory is divided as equally as possible among any partitions with a zero or blank size parameter. Dividing the remaining memory is subject to the following constraints:

- Blocks must be allocated in multiples of 4 KB.
- Addressing constraints might leave un-allocated areas of memory between partitions.

See the `modprobe` man page for details about `modprobe`.

To load the XPRAM device driver module automatically at boot time, see the section on persistent module loading in `Red Hat Enterprise Linux 7.5 System Administrator’s Guide`.

**Examples**

- The following specification allocates the extended storage into four partitions. Partition 1 has 2 GB (2097152 KB), partition 4 has 4 GB (4194304 KB), and partitions 2 and 3 use equal parts of the remaining storage. If the total amount of extended storage was 16 GB, then partitions 3 and 4 would each have approximately 5 GB.

  ```bash
  # modprobe xpram devs=4 sizes=2097152,0,0,4194304
  ```

- The following specification allocates the extended storage into three partitions. The partition 2 has 512 KB and the partitions 1 and 3 use equal parts of the remaining extended storage.

  ```bash
  # modprobe xpram devs=3 sizes=512
  ```

- The following specification allocates the extended storage into two partitions of equal size.

  ```bash
  # modprobe xpram devs=2
  ```
Part 4. Networking

Chapter 15. qeth device driver for OSA-Express (QDIO) and HiperSockets ........................................... 223
   Device driver functions ........................................... 226
   What you should know about the qeth device driver .......... 229
   Setting up the qeth device driver ................................. 237
   Working with qeth devices ......................................... 238
   Working with qeth devices in layer 3 mode .................... 259
   Working with qeth devices in layer 2 mode .................... 269
   Scenario: VIPA – minimize outage due to adapter failure ... 272
   Scenario: Virtual LAN (VLAN) support ........................ 277
   HiperSockets Network Concentrator .............................. 280
   Setting up for DHCP with IPv4 .................................. 285
   Setting up Linux as a LAN sniffer ............................... 286

Chapter 16. OSA-Express SNMP subagent support ........................................... 289
   What you should know about osasnmpd ......................... 289
   Setting up osasnmpd .................................................. 290
   Working with the osasnmpd subagent ............................ 294

Chapter 17. LAN channel station device driver ........................................... 299
   What you should know about LCS ................................. 299
   Setting up the LCS device driver ................................. 300
   Working with LCS devices .......................................... 300

Chapter 18. CTCM device driver ........................................... 305
   Features ................................................................. 305
   What you should know about CTCM ............................ 305
   Setting up the CTCM device driver .............................. 307
   Working with CTCM devices ........................................ 307
   CTCM scenarios ........................................................ 314

Chapter 19. AF_IUCV address family support ........................................... 319
   Features ................................................................. 319
   Setting up the AF_IUCV address family support ................ 320
   Addressing AF_IUCV sockets in applications .................... 321

Chapter 20. RDMA over Converged Ethernet ........................................... 323
   Working with the RoCE support .................................... 323
   Enabling debugging .................................................. 324

Red Hat Enterprise Linux 7.5 includes several network device drivers that are specific to z/Architecture.

Newest version

You can find the newest version of this book at
www.ibm.com/support/knowledgecenter/linuxonibm/linz_r_distlibs.html

Restrictions

For prerequisites and restrictions see the IBM Z architecture specific information in the Red Hat Enterprise Linux 7.5 release notes at
https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux

Example

An example network setup that uses some available network setup types is shown in Figure 43 on page 220
In the example there are three Linux instances; two of them run as z/VM guests in one LPAR and a third Linux instance runs in another LPAR. Within z/VM, Linux instances can be connected through a guest LAN or VSWITCH. Within and between LPARs, you can connect Linux instances through HiperSockets. OSA-Express cards running in either non-QDIO mode (called LCS here) or in QDIO mode can connect the mainframe to an external network.

Figure 43. Networking example
Table 33 lists which control units and device type combinations are supported by the network device drivers.

**Table 33. Supported device types, control units, and corresponding device drivers**

<table>
<thead>
<tr>
<th>Device type</th>
<th>Control unit</th>
<th>Device driver</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1732/01</td>
<td>1731/01</td>
<td>qeth</td>
<td>OSA configured as OSD</td>
</tr>
<tr>
<td>1732/02</td>
<td>1731/02</td>
<td>qeth</td>
<td>OSA configured as OSX</td>
</tr>
<tr>
<td>1732/03</td>
<td>1731/02</td>
<td>qeth</td>
<td>OSA configured as OSM</td>
</tr>
<tr>
<td>1732/05</td>
<td>1731/05</td>
<td>qeth</td>
<td>HiperSockets</td>
</tr>
<tr>
<td>1732/06</td>
<td>1731/06</td>
<td>qeth</td>
<td>OSA configured as OSN</td>
</tr>
<tr>
<td>0000/00</td>
<td>3088/01</td>
<td>lcs</td>
<td>P/390</td>
</tr>
<tr>
<td>0000/00</td>
<td>3088/08</td>
<td>ctcm</td>
<td>Virtual CTC under z/VM</td>
</tr>
<tr>
<td>0000/00</td>
<td>3088/1e</td>
<td>ctcm</td>
<td>FICON channel</td>
</tr>
<tr>
<td>0000/00</td>
<td>3088/1f</td>
<td>lcs</td>
<td>2216 Nways Multiaccess Connector</td>
</tr>
<tr>
<td>0000/00</td>
<td>3088/1f</td>
<td>ctcm</td>
<td>ESCON channel</td>
</tr>
<tr>
<td>0000/00</td>
<td>3088/60</td>
<td>lcs</td>
<td>OSA configured as OSE (non-QDIO)</td>
</tr>
</tbody>
</table>
Chapter 15. qeth device driver for OSA-Express (QDIO) and HiperSockets

The qeth device driver supports a multitude of network connections, for example, connections through Open Systems Adapters (OSA), HiperSockets, guest LANs, and virtual switches.

Real connections that use OSA-Express

An IBM mainframe uses OSA-Express adapters, which are real LAN-adapter hardware, see [Figure 44]. These adapters provide connections to the outside world, but can also connect virtual systems (between LPARs or between z/VM guest virtual machines) within the mainframe. The qeth driver supports these adapters if they are defined to run in queued direct I/O (QDIO) mode (defined as OSD in the hardware configuration). OSD-devices are the standard IBM Z LAN-converters. For details about OSA-Express in QDIO mode, see Open Systems Adapter-Express Customer’s Guide and Reference, SA22-7935.

Deprecated connection type: OSN devices are deprecated. Do not use for new network setups.

Figure 44. OSA-Express adapters are real LAN-adapter hardware

The qeth device driver supports OSA-Express features for the IBM Z mainframes that are relevant to Red Hat Enterprise Linux 7.5 as shown in Table 34 on page 224.
Table 34: The qeth device driver support for OSA-Express features

<table>
<thead>
<tr>
<th>Feature</th>
<th>z14</th>
<th>z13 and z13s</th>
<th>zEC12 and zBC12</th>
<th>z196 and z114</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSA-Express6S</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
<td>Not supported</td>
<td>Not supported</td>
</tr>
<tr>
<td></td>
<td>1000Base-T Ethernet</td>
<td>1000Base-T Ethernet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSA-Express5S</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
<td>Not supported</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000Base-T Ethernet</td>
<td>1000Base-T Ethernet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSA-Express4S</td>
<td>Not supported</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000Base-T Ethernet</td>
<td>1000Base-T Ethernet</td>
<td></td>
</tr>
<tr>
<td>OSA-Express3</td>
<td>Not supported</td>
<td>Not supported</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
<td>Gigabit Ethernet 10 Gigabit Ethernet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000Base-T Ethernet</td>
<td>1000Base-T Ethernet</td>
<td></td>
</tr>
<tr>
<td>OSA-Express2</td>
<td>Not supported</td>
<td>Not supported</td>
<td>Not supported</td>
<td>Gigabit Ethernet 1000Base-T Ethernet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000Base-T Ethernet</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Unless otherwise indicated, OSA-Express refers to the OSA-express features as shown in Table 34.

The qeth device driver supports CHPIDs of type OSM and OSX:

OSM provides connectivity to the intranode management network (INMN) from Unified Resource Manager functions to a zEnterprise CPC.

OSX provides connectivity to and access control for the intraensemble data network (IEDN), which is managed by Unified Resource Manager functions. A zEnterprise CPC and zBX within an ensemble are connected through the IEDN. See zEnterprise System Introduction to Ensembles, GC27-2609 and zEnterprise System Ensemble Planning and Configuring Guide, GC27-2608 for more details.

HiperSockets
An IBM mainframe uses internal connections that are called HiperSockets. These simulate QDIO network adapters and provide high-speed TCP/IP communication for operating system instances within and across LPARs. For details about HiperSockets, see HiperSockets Implementation Guide, SG24-6816.

Virtual connections for Linux on z/VM
z/VM offers virtualized LAN-adapters that enable connections between z/VM guest virtual machines and the outside world. It allows definitions of simulated network interface cards (NICs) attached to certain z/VM guest virtual machines. The NICs can be connected to a simulated LAN segment called guest LAN for z/VM internal communication between z/VM guest virtual machines, or they can be connected to a virtual switch called VSWITCH for external LAN connectivity.

Guest LAN
Guest LANs represent a simulated LAN segment that can be connected to simulated network interface cards. There are three types of guest LANs:
- Simulated OSA-Express in layer 3 mode
- Simulated HiperSockets (layer 3) mode
Simulated Ethernet in layer 2 mode

Each guest LAN is isolated from other guest LANs on the same system (unless some member of one LAN group acts as a router to other groups). See Figure 45.

Virtual switch

A virtual switch (VSWITCH) is a special-purpose guest LAN that provides external LAN connectivity through an additional OSA-Express device served by z/VM without the need for a routing virtual machine, see Figure 46 on page 226.
A dedicated OSA adapter can be an option, but is not required for a VSWITCH.

From a Linux point of view there is no difference between guest LAN- and VSWITCH-devices; thus Linux talks about guest LAN-devices independently of their z/VM-attachment to a guest LAN or VSWITCH.

For information about guest LANs, virtual switches, and virtual HiperSockets, see z/VM Connectivity, SC24-6174.

### Device driver functions

The qeth device driver supports many networking transport protocol functions, as well as offload functions and problem determination functions.

The qeth device driver supports functions listed in Table 35 and Table 36 on page 228.

#### Table 35. Real connections

<table>
<thead>
<tr>
<th>Function</th>
<th>OSA Layer 2</th>
<th>OSA Layer 3</th>
<th>HiperSockets Layer 2 Ethernet</th>
<th>HiperSockets Layer 3 Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic device or protocol functions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPv4/multicast/broadcast</td>
<td>Yes/Yes/Yes</td>
<td>Yes/Yes/Yes</td>
<td>Yes/Yes/Yes</td>
<td>Yes/Yes/Yes</td>
</tr>
<tr>
<td>IPv6/multicast</td>
<td>Yes/Yes</td>
<td>Yes/Yes</td>
<td>Yes/Yes</td>
<td>Yes/Yes</td>
</tr>
<tr>
<td>Non-IP traffic</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>VLAN IPv4/IPv6/non IP</td>
<td>sw/sw/sw</td>
<td>hw/sw/sw</td>
<td>sw/sw/sw</td>
<td>hw/sw/No</td>
</tr>
<tr>
<td>Linux ARP</td>
<td>Yes</td>
<td>No</td>
<td>(hw ARP)</td>
<td>Yes</td>
</tr>
<tr>
<td>Linux neighbor solicitation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Unique MAC address</td>
<td>Yes (random)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Table 35. Real connections (continued)

<table>
<thead>
<tr>
<th>Function</th>
<th>OSA Layer 2</th>
<th>OSA Layer 3</th>
<th>HiperSockets Layer 2 Ethernet</th>
<th>HiperSockets Layer 3 Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change MAC address</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Promiscuous mode</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>• Yes (for sniffer=1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• No (for sniffer=0)</td>
</tr>
<tr>
<td>MAC headers send/receive</td>
<td>Yes/Yes</td>
<td>faked/faked</td>
<td>Yes/Yes</td>
<td>faked/faked</td>
</tr>
<tr>
<td>ethtool support</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Bonding</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Priority queueing</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Secondary unicast MAC address</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Bridge port</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

**Offload features**

<table>
<thead>
<tr>
<th>Function</th>
<th>OSA Layer 2</th>
<th>OSA Layer 3</th>
<th>HiperSockets Layer 2 Ethernet</th>
<th>HiperSockets Layer 3 Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP segmentation offload (TSO)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Inbound (rx) checksum</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Outbound (tx) checksum</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

**OSA/QETH specific features**

<table>
<thead>
<tr>
<th>Function</th>
<th>OSA Layer 2</th>
<th>OSA Layer 3</th>
<th>HiperSockets Layer 2 Ethernet</th>
<th>HiperSockets Layer 3 Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Special device driver setup for VIPA</td>
<td>No</td>
<td>required</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Special device driver setup for proxy ARP</td>
<td>No</td>
<td>required</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Special device driver setup for IP takeover</td>
<td>No</td>
<td>required</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Special device driver setup for routing IPv4/IPv6</td>
<td>No/No</td>
<td>required/required</td>
<td>No/No</td>
<td>Yes/Yes</td>
</tr>
<tr>
<td>Receive buffer count</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Direct connectivity to z/OS</td>
<td>Yes by HW</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>SNMP support</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Multiport support</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Data connection isolation</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

**Problem determination**

<table>
<thead>
<tr>
<th>Function</th>
<th>OSA Layer 2</th>
<th>OSA Layer 3</th>
<th>HiperSockets Layer 2 Ethernet</th>
<th>HiperSockets Layer 3 Ethernet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware trace</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

**Legend:**

- **No** Function not supported or not required.
- **Yes** Function supported.
- **hw** Function performed by hardware.
- **sw** Function performed by software.
- **faked** Function will be simulated.
- **required** Function requires special setup.
<table>
<thead>
<tr>
<th>Function</th>
<th>Emulated OSA Layer 2</th>
<th>Emulated OSA Layer 3</th>
<th>Emulated HiperSockets Layer 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic device or protocol features</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPv4/multicast/broadcast</td>
<td>Yes/Yes/Yes</td>
<td>Yes/Yes/Yes</td>
<td>Yes/Yes/Yes</td>
</tr>
<tr>
<td>IPv6/multicast</td>
<td>Yes/Yes</td>
<td>Yes/Yes</td>
<td>No/No</td>
</tr>
<tr>
<td>Non-IP traffic</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>VLAN IPv4/IPv6/non IP</td>
<td>sw/sw/sw</td>
<td>hw/sw/No</td>
<td>hw/No/No</td>
</tr>
<tr>
<td>Linux ARP</td>
<td>Yes</td>
<td>No (hw ARP)</td>
<td>No</td>
</tr>
<tr>
<td>Linux neighbor solicitation</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Unique MAC address</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Change MAC address</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Promiscuous mode</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>MAC headers send/receive</td>
<td>Yes/Yes</td>
<td>faked/faked</td>
<td>faked/faked</td>
</tr>
<tr>
<td>Ethtool support</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Bonding</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Priority queueing</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Secondary unicast MAC address</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Offload features</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>OSA/QETH specific features</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Special device driver setup for VIPA</td>
<td>No</td>
<td>required</td>
<td>required</td>
</tr>
<tr>
<td>Special device driver setup for proxy ARP</td>
<td>No</td>
<td>required</td>
<td>required</td>
</tr>
<tr>
<td>Special device driver setup for IP takeover</td>
<td>No</td>
<td>required</td>
<td>required</td>
</tr>
<tr>
<td>Special device driver setup for routing IPv4/IPv6</td>
<td>No/No</td>
<td>required/required</td>
<td>required/required</td>
</tr>
<tr>
<td>Receive buffer count</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Direct connectivity to z/OS</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SNMP support</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Multiport support</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Data connection isolation</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Problem determination</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hardware trace</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Legend:
- No Function not supported or not required.
- Yes Function supported.
- hw Function performed by hardware.
- sw Function performed by software.
- faked Function will be simulated.
- required Function requires special setup.
What you should know about the qeth device driver

Interface names are assigned to qeth group devices, which map to subchannels and their corresponding device numbers and device bus-IDs. An OSA-Express adapter can handle both IPv4 and IPv6 packets.

Layer 2 and layer 3

The qeth device driver consists of a common core and two device disciplines: layer 2 and layer 3.

In layer 2 mode, OSA routing to the destination Linux instance is based on MAC addresses. A local MAC address is assigned to each interface of a Linux instance and registered in the OSA Address Table. These MAC addresses are unique and different from the MAC address of the OSA adapter. See “MAC headers in layer 2 mode” on page 231 for details.

In layer 3 mode, all interfaces of all Linux instances share the MAC address of the OSA adapter. OSA routing to the destination Linux instance is based on IP addresses. See “MAC headers in layer 3 mode” on page 233 for details.

The layer 2 discipline (qeth_l2)

The layer 2 discipline supports:
- OSA devices and z/VM virtual NICs that couple to VSWITCHes or QDIO guest LANs
- OSA for NCP devices
- HiperSockets devices
- OSM (OSA-Express for Unified Resource Manager) devices
- OSX (OSA-Express for zBX) devices for IEDN

The layer 2 discipline is the default setup for OSA. On HiperSockets the default continues to be layer 3. OSA guest LANs are layer 2 by default, while HiperSockets guest LANs are always layer 3. See “Setting the layer2 attribute” on page 243 for details.

The layer 3 discipline (qeth_l3)

The layer 3 discipline supports:
- OSA devices and z/VM virtual NICs that couple to VSWITCHes or QDIO guest LANs that are running in layer 3 mode (with faked link layer headers)
- HiperSockets and HiperSockets guest LAN devices that are running in layer 3 mode (with faked link layer headers)
- OSX (OSA-Express for zBX) devices for IEDN

This discipline supports those devices that are not capable of running in layer 2 mode. Not all Linux networking features are supported and others need special setup or configuration. See Table 41 on page 240. Some performance-critical applications might benefit from being layer 3.

Layer 2 and layer 3 interfaces cannot communicate within a HiperSockets LAN or within a VSWITCH or guest LAN. However, a shared OSA adapter can convert traffic between layer 2 and layer 3 networks.
qeth group devices

The qeth device driver requires three I/O subchannels for each HiperSockets CHPID or OSA-Express CHPID in QDIO mode. One subchannel is for control reads, one for control writes, and the third is for data.

The qeth device driver uses the QDIO protocol to communicate with the HiperSockets and OSA-Express adapter.

![Figure 47. I/O subchannel interface](image)

The three device bus-IDs that correspond to the subchannel triplet are grouped as one qeth group device. The following rules apply for the device bus-IDs:

- **read** no specific rules.
- **write** must be the device bus-ID of the read subchannel plus one.
- **data** can be any free device bus-ID on the same CHPID.

You can configure different triplets of device bus-IDs on the same CHPID differently. For example, if you have two triplets on the same CHPID they can have different attribute values for priority queueing.

Overview of the steps for setting up a qeth group device

You need to perform several steps before user-space applications on your Linux instance can use a qeth group device.

**Before you begin**

Find out how the hardware is configured and which qeth device bus-IDs are on which CHPID, for example by looking at the IOCDS. Identify the device bus-IDs that you want to group into a qeth group device. The three device bus-IDs must be on the same CHPID.

**Procedure**

Perform these steps to allow user-space applications on your Linux instance to use a qeth group device:

1. Create the qeth group device.
   
   After booting Linux, each qeth device bus-ID is represented by a subdirectory in `/sys/bus/ccw/devices/`. These subdirectories are then named with the bus IDs of the devices. For example, a qeth device with bus IDs 0.0.fc00, 0.0.fc01, and 0.0.fc02 is represented as `/sys/bus/ccw/drivers/qeth/0.0.fc00`

2. Configure the device.

3. Set the device online.
4. Activate the device and assign an IP address to it.

**What to do next**

These tasks and the configuration options are described in detail in “Working with qeth devices” on page 238.

**qeth interface names and device directories**

Red Hat Enterprise Linux automatically assigns interface names to the qeth group devices. The qeth device driver creates the corresponding sysfs structures.

While an interface is online, it is represented in sysfs as:

/sys/class/net/<interface>

The mapping between interface names and the device bus-ID that represents the qeth group device in sysfs is preserved when a device is set offline and back online.

“Finding out the interface name of a qeth group device” on page 248 and “Finding out the bus ID of a qeth interface” on page 249 provide information about mapping device bus-IDs and interface names.

**Support for IP Version 6 (IPv6)**

The qeth device driver supports IPv6 in many network setups.

IPv6 is supported on:

- Ethernet interfaces of the OSA-Express adapter that runs in QDIO mode.
- HiperSockets layer 2 and layer 3 interfaces.
- z/VM guest LAN interfaces running in QDIO or HiperSockets layer 3 mode.
- z/VM guest LAN and VSWITCH interfaces in layer 2.

There are noticeable differences between the IP stacks for versions 4 and 6. Some concepts in IPv6 are different from IPv4, such as neighbor discovery, broadcast, and Internet Protocol security (IPsec). IPv6 uses a 16-byte address field, while the addresses under IPv4 are 4 bytes in length.

Stateless autoconfiguration generates unique IP addresses for all Linux instances, even if they share an OSA-Express adapter with other operating systems.

Be aware of the IP version when you specify IP addresses and when you use commands that return IP version-specific output (such as `qetharp`).

**MAC headers in layer 2 mode**

In LAN environments, data packets find their destination through Media Access Control (MAC) addresses in their MAC header.
MAC address handling as shown in Figure 48 applies to non-mainframe environments and a mainframe environment with an OSA-Express adapter where the layer2 option is enabled.

The layer2 option keeps the MAC addresses on incoming packets. Incoming and outgoing packets are complete with a MAC header at all stages between the Linux network stack and the LAN as shown in Figure 48. This layer2-based forwarding requires unique MAC addresses for all concerned Linux instances.

In layer 2 mode, the Linux TCP/IP stack has full control over the MAC headers and the neighbor lookup. The Linux TCP/IP stack does not configure IPv4 or IPv6 addresses into the hardware, but requires a unique MAC address for the card.

When Linux is running as a z/VM guest, the qeth device driver obtains a MAC address for each L2 device from the z/VM host. No configuration is necessary.

When Linux is running in an LPAR and you work with a directly attached OSA adapter in QDIO mode, you should assign a unique MAC address.

To assign a MAC address, add a line MACADDR='MAC address' to the configuration file /etc/sysconfig/network-scripts/ifcfg-*<if-name>*. Alternatively, you can set the MAC address by issuing the command:

```
ip link set addr <MAC address> dev <interface>
```

**Note:** Be sure not to assign the MAC address of the OSA-Express adapter to your Linux instance.

For OSX and OSM CHPIDs, you cannot set your own MAC addresses. Linux uses the MAC addresses defined by the Unified Resource Manager.

For HiperSockets connections, a MAC address is generated. For most purposes the generated address is adequate. However, you can change the address by using the `ip` command if you need to.

For connections within a QDIO-based z/VM guest LAN environment, z/VM assigns the necessary MAC addresses to its guests.
MAC headers in layer 3 mode

A qeth layer 3 mode device driver is an Ethernet offload engine for IPv4 and a partial Ethernet offload engine for IPv6. Hence, there are some special things to understand about the layer 3 mode.

To support IPv6 and protocols other than IPv4, the device driver registers a layer 3 card as an Ethernet device to the Linux TCP/IP stack.

In layer 3 mode, the OSA-Express adapter in QDIO mode removes the MAC header with the MAC address from incoming IPv4 packets. It uses the registered IP addresses to forward a packet to the recipient TCP/IP stack. See Figure 49. Thus the OSA-Express adapter is able to deliver IPv4 packets to the correct Linux images. Apart from broadcast packets, a Linux image can get only packets for IP addresses it configured in the stack and registered with the OSA-Express adapter.

Figure 49. MAC address handling in layer3 mode

The OSA-Express QDIO microcode builds MAC headers for outgoing IPv4 packets and removes them from incoming IPv4 packets. Hence, the operating systems' network stacks send and receive only IPv4 packets without MAC headers.

This lack of MAC headers can be a problem for applications that expect MAC headers. For examples of how such problems can be resolved, see "Setting up for DHCP with IPv4" on page 285.

Outgoing frames

The qeth device driver registers the layer 3 card as an Ethernet device. Therefore, the Linux TCP/IP stack will provide complete Ethernet frames to the device driver.

If the hardware does not require the Ethernet frame (for example, for IPv4) the driver removes the Ethernet header prior to sending the frame to the hardware. If necessary information like the Ethernet target address is not available (because of the offload functionality) the value is filled with the hardcoded address FAKELL.
Table 37. Ethernet addresses of outgoing frames

<table>
<thead>
<tr>
<th>Frame</th>
<th>Destination address</th>
<th>Source address</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPv4</td>
<td>FAKELL</td>
<td>Real device address</td>
</tr>
<tr>
<td>IPv6</td>
<td>Real destination address</td>
<td>Real device address</td>
</tr>
<tr>
<td>Other packets</td>
<td>Real destination address</td>
<td>Real device address</td>
</tr>
</tbody>
</table>

Incoming frames
The device driver provides Ethernet headers for all incoming frames.

If necessary information like the Ethernet source address is not available (because of the offload functionality) the value is filled with the hardcoded address FAKELL.

Table 38. Ethernet addresses of incoming frames

<table>
<thead>
<tr>
<th>Frame</th>
<th>Destination address</th>
<th>Source address</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPv4</td>
<td>Real device address</td>
<td>FAKELL</td>
</tr>
<tr>
<td>IPv6</td>
<td>Real device address</td>
<td>FAKELL</td>
</tr>
<tr>
<td>Other packets</td>
<td>Real device address</td>
<td>Real source address</td>
</tr>
</tbody>
</table>

Note that if a source or destination address is a multicast or broadcast address the device driver can provide the corresponding (real) Ethernet multicast or broadcast address even when the packet was delivered or sent through the offload engine. Always providing the link layer headers enables packet socket applications like tcpdump to work properly on a qeth layer 3 device without any changes in the application itself (the patch for libpcap is no longer required).

While the faked headers are syntactically correct, the addresses are not authentic, and hence applications requiring authentic addresses will not work. Some examples are given in Table 39.

Table 39. Applications that react differently to faked headers

<table>
<thead>
<tr>
<th>Application</th>
<th>Support</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>tcpdump</td>
<td>Yes</td>
<td>Displays only frames, fake Ethernet information is displayed.</td>
</tr>
<tr>
<td>iptables</td>
<td>Partially</td>
<td>As long as the rule does not deal with Ethernet information of an IPv4 frame.</td>
</tr>
<tr>
<td>dhcp</td>
<td>Yes</td>
<td>Is non-IPv4 traffic.</td>
</tr>
</tbody>
</table>

IP addresses
The network stack of each operating system that shares an OSA-Express adapter in QDIO mode registers all its IP addresses with the adapter.

Whenever IP addresses are deleted from or added to a network stack, the device drivers download the resulting IP address list changes to the OSA-Express adapter.

For the registered IP addresses, the OSA-Express adapter off-loads various functions, in particular also:

- Handling MAC addresses and MAC headers
- ARP processing
ARP:

The OSA-Express adapter in QDIO mode responds to Address Resolution Protocol (ARP) requests for all registered IPv4 addresses.

ARP is a TCP/IP protocol that translates 32-bit IPv4 addresses into the corresponding hardware addresses. For example, for an Ethernet device, the hardware addresses are 48-bit Ethernet Media Access Control (MAC) addresses. The mapping of IPv4 addresses to the corresponding hardware addresses is defined in the ARP cache. When it needs to send a packet, a host consults the ARP cache of its network adapter to find the MAC address of the target host.

If there is an entry for the destination IPv4 address, the corresponding MAC address is copied into the MAC header and the packet is added to the appropriate interface’s output queue. If the entry is not found, the ARP functions retain the IPv4 packet, and broadcast an ARP request asking the destination host for its MAC address. When a reply is received, the packet is sent to its destination.

Notes:

1. On an OSA-Express adapter in QDIO mode, do not set the NO_ARP flag on the Linux Ethernet device. The device driver disables the ARP resolution for IPv4. Because the hardware requires no neighbor lookup for IPv4, but neighbor solicitation for IPv6, the NO_ARP flag is not allowed on the Linux Ethernet device.

2. On HiperSockets, which is a full Ethernet offload engine for IPv4 and IPv6 and supports no other traffic, the device driver sets the NO_ARP flag on the Linux Ethernet interface. Do not remove this flag from the interface.

Layer 2 bridge port function

OSA and HiperSockets ports that operate in layer 2 mode can be set up to receive all frames that are addressed to unknown MAC addresses.

Other architectures

Non IBM Z networks use Ethernet Network Interface Controllers (NICs) to pass traffic between the operating system and the network. Normally, a NIC filters incoming traffic to admit only frames with destination MAC addresses that match addresses that are registered with the NIC.

However, a NIC can also be configured to receive and pass to the operating system all Ethernet frames that reach it, regardless of the destination MAC address. This mode of operation is known as “promiscuous mode”. For example, promiscuous mode is a prerequisite for configuring a NIC as a member of a Linux software bridge.

For more information about how to set up a software bridge, see the documentation that is provided by Red Hat Enterprise Linux, or the bridging how-to available at [http://www.tldp.org/HOWTO/BRIDGE-STP-HOWTO](http://www.tldp.org/HOWTO/BRIDGE-STP-HOWTO)

z/Architecture

OSA and HiperSockets adapters on IBM Z do not have a direct equivalent of promiscuous mode of operation. Instead, OSA and HiperSockets hardware support
bridge port functions. The operating system can assign a bridge port role to a logical port, and the adapter assigns an active state to one of the logical ports to which a role was assigned.

A local port in active bridge port state receives all Ethernet frames with unknown destination MAC addresses. [Figure 50] shows a setup with a HiperSockets bridge port and an OSA bridge port.

![Figure 50. HiperSockets and OSA bridge port in Linux](image)

HiperSockets only: Permission to configure ports as bridge ports must be granted in IBM zEnterprise Unified Resource Manager (zManager).

**Differences between promiscuous mode and bridge-port roles**

Making a logical port of an OSA or HiperSockets adapter an active bridge port is similar to enabling promiscuous mode on a non-mainframe NIC that is connected to a real Ethernet switch. However, there are important differences:

**Number of ports in promiscuous mode**

- Real switches: Any number of interfaces that are connected to a real switch can be turned to promiscuous mode, and all of them then receive frames with unknown destination addresses.
- Bridge ports on Z: Although you can assign the bridge-port role to multiple ports of a single OSA or HiperSockets adapter, only one port is active and receives traffic to unknown destinations.

**Interception of traffic to other systems**

- Real switches: A port of a real switch can be configured to receive frames with both known and unknown destinations. If a NIC in promiscuous mode is connected to the port, the corresponding host receives a copy of all traffic that passes through the switch. This includes traffic that is destined to other hosts connected to this switch.
• Bridge ports on Z: Only frames with unknown destinations are passed to the operating system. It is not possible to intercept traffic addressed to systems connected to other ports of the same OSA adapter.

**Limitation by the source of traffic**

• Real switches and HiperSockets bridge-port LAN: Frames with unknown destination MAC addresses are delivered to the promiscuous interfaces regardless of the port through which the frames enter the switch or HiperSockets adapter.

• OSA bridge port only: An active bridge port learns which MAC addresses need to be routed to the owning system by analyzing ARP and other traffic. Incoming frames are routed to the active bridge port if their destination MAC address:
  - Matches an address that is learned or registered with the bridge port
  - Is not learned or registered with any of the local ports of the OSA adapter, but arrived from the physical Ethernet port

**Bridge port roles**

Linux can assign a primary or secondary role to a logical port of an OSA or a HiperSockets adapter. Only one logical port of such an adapter can be assigned the primary role, but multiple other logical ports can be assigned secondary role. When one or more logical ports of an adapter are assigned primary or secondary role, the hardware ensures that exactly one of these ports is active. The active port receives frames with unknown destination. When a port with primary role is present, it always becomes active. When only ports with secondary role are present, the hardware decides which one becomes active. Changes in the ports' state are reported to Linux user space through udev events.

You can set a bridge port role either directly by using the `bridge_role` attribute or indirectly by using the `bridge_reflect_promisc` attribute. See “Configuring a network device as a member of a Linux bridge” on page 269.

---

**Setting up the qeth device driver**

No module parameters exist for the qeth device driver. qeth devices are set up using sysfs.

**Loading the qeth device driver modules**

You must load the qeth device driver before you can work with qeth devices.

Use the `modprobe` command to load the qeth device driver, and to automatically load all required additional modules in the correct order:

```
modprobe qeth qeth_l2 qeth_l3
```

---

Chapter 15. qeth: OSA-Express (QDIO) and HiperSockets 237
where:

- qeth is the core module that contains common functions that are used for both layer 2 and layer 3 disciplines.
- qeth_l2 is the module that contains layer 2 discipline-specific code.
- qeth_l3 is the module that contains layer 3 discipline-specific code.

When a qeth device is configured for a particular discipline, the driver tries to automatically load the corresponding discipline module.

**Switching the discipline of a qeth device**

To switch the discipline of a device, the network interface must be shut down and the device must be offline.

Some devices can only run in one discipline, see “Layer 2 and layer 3” on page 229. The device driver rejects any request to switch the discipline of these devices.

If the new discipline is accepted by the device driver, the old network interface is deleted. When the new discipline is set online the first time, the new network interface is created.

**Removing the modules**

Removing a module is not possible if there are cross dependencies between the discipline modules and the core module.

To release the dependencies from the core module to the discipline module, all devices of this discipline must be ungrouped. Now the discipline module can be removed. If all discipline modules are removed, the core module can be removed.

**Working with qeth devices**

Typical tasks that you need to perform when working with qeth devices include creating group devices, finding out the type of a network adapter, and setting a device online or offline.

**About this task**

Most of these tasks involve writing to and reading from attributes of qeth group devices in sysfs. This is useful on a running system where you want to make dynamic changes. If you want to make the changes persistent across IPLs, use the interface configuration files. Network configuration parameters are defined in /etc/sysconfig/network-scripts/ifcfg-*<if_name>*. An example of how to define a qeth device persistently is in Red Hat Enterprise Linux 7.5 Installation Guide. For a general discussion of network configuration files, see Red Hat Enterprise Linux 7.5 System Administrator’s Guide.

Table 40 on page 239 and Table 41 on page 240 serve as both a task overview and a summary of the attributes and the possible values you can write to them. Underlined values are defaults.
Tip: Use the `znetconf` command to configure devices instead of using the attributes directly (see “znetconf - List and configure network devices” on page 648).

Not all attributes are applicable to each device. Some attributes apply only to HiperSockets or only to OSA-Express CHPIPs in QDIO mode, other attributes are applicable to IPv4 interfaces only. See the task descriptions for the applicability of each attribute.

OSA for NCP handles NCP-related packets. Most of the attributes do not apply to OSA for NCP devices. The attributes that apply are:

- `if_name`
- `card_type`
- `buffer_count`
- `recover`

Table 40. qeth tasks and attributes common to layer2 and layer3

<table>
<thead>
<tr>
<th>Task</th>
<th>Corresponding attributes</th>
<th>Possible attribute values</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Creating a qeth group device” on page 241</td>
<td>group</td>
<td>n/a</td>
</tr>
<tr>
<td>“Removing a qeth group device” on page 242</td>
<td>ungroup</td>
<td>0 or 1</td>
</tr>
<tr>
<td>“Setting the layer2 attribute” on page 243</td>
<td>layer2</td>
<td>0 or 1, see “Layer 2 and layer 3” on page 229</td>
</tr>
<tr>
<td>“Using priority queueing” on page 244</td>
<td>priority_queueing</td>
<td>prio_queueing_vlan, prio_queueing_skb, prio_queueing_prec, prio_queueing/tos, no_prio_queueing, no_prio_queueing:0, no_prio_queueing:1, no_prio_queueing:2, no_prio_queueing:3</td>
</tr>
<tr>
<td>“Specifying the number of inbound buffers” on page 246</td>
<td>buffer_count</td>
<td>integer in the range 8 - 128. The default is 64 for OSA devices and 128 for HiperSockets devices</td>
</tr>
<tr>
<td>“Specifying the relative port number” on page 246</td>
<td>portno</td>
<td>integer, either 0 or 1, the default is 0</td>
</tr>
<tr>
<td>“Finding out the type of your network adapter” on page 247</td>
<td>card_type</td>
<td>n/a, read-only</td>
</tr>
<tr>
<td>“Setting a device online or offline” on page 248</td>
<td>online</td>
<td>0 or 1</td>
</tr>
<tr>
<td>“Finding out the interface name of a qeth group device” on page 248</td>
<td>if_name</td>
<td>n/a, read-only</td>
</tr>
<tr>
<td>“Finding out the bus ID of a qeth interface” on page 249</td>
<td>none</td>
<td>n/a</td>
</tr>
<tr>
<td>“Activating an interface” on page 249</td>
<td>none</td>
<td>n/a</td>
</tr>
<tr>
<td>“Deactivating an interface” on page 251</td>
<td>none</td>
<td>n/a</td>
</tr>
<tr>
<td>“Recovering a device” on page 252</td>
<td>recover</td>
<td>1</td>
</tr>
<tr>
<td>“Turning inbound checksum calculations on and off” on page 253</td>
<td>none</td>
<td>n/a</td>
</tr>
<tr>
<td>“Turning outbound checksum calculations on and off” on page 253</td>
<td>none</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Table 40. qeth tasks and attributes common to layer2 and layer3 (continued)

<table>
<thead>
<tr>
<th>Task</th>
<th>Corresponding attributes</th>
<th>Possible attribute values</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Isolating data connections” on page 254</td>
<td>isolation</td>
<td>none, drop, forward</td>
</tr>
<tr>
<td>“Starting and stopping collection of QETH performance statistics” on page 257</td>
<td>performance_stats</td>
<td>0 or 1</td>
</tr>
<tr>
<td>“Capturing a hardware trace” on page 258</td>
<td>hw_trap</td>
<td>arm, disarm</td>
</tr>
</tbody>
</table>

¹A value of -1 means that the layer has not been set and that the default layer setting is used when the device is set online.

Table 41. qeth tasks and attributes in layer 3 mode

<table>
<thead>
<tr>
<th>Task</th>
<th>Corresponding attributes</th>
<th>Possible attribute values</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Setting up a Linux router in layer 3” on page 260</td>
<td>route4, route6</td>
<td>primary_router, secondary_router, primary_connector, secondary_connector, multicast_router, no_router</td>
</tr>
<tr>
<td>“Enabling and disabling TCP segmentation offload” on page 262</td>
<td>none</td>
<td>n/a</td>
</tr>
<tr>
<td>“Faking broadcast capability” on page 263</td>
<td>fake_broadcast ¹</td>
<td>0 or 1</td>
</tr>
<tr>
<td>“Taking over IP addresses” on page 263</td>
<td>ipa_takeover/enable, ipa_takeover/add4, ipa_takeover/add6, ipa_takeover/del4, ipa_takeover/del6, ipa_takeover/invert4, ipa_takeover/invert6</td>
<td>IPv4 or IPv6 IP address and mask bits</td>
</tr>
<tr>
<td>“Configuring a device for proxy ARP” on page 267</td>
<td>rxip/add4, rxip/del4</td>
<td>IPv4 IP address</td>
</tr>
<tr>
<td>Configuring a device for NDP proxy</td>
<td>rxip/add6, rxip/del6</td>
<td>IPv6 IP address</td>
</tr>
<tr>
<td>“Configuring a device for virtual IP address (VIPA)” on page 268</td>
<td>vipa/add4, vipa/add6, vipa/del4, vipa/del6</td>
<td>IPv4 or IPv6 IP address</td>
</tr>
<tr>
<td>“Configuring a HiperSockets device for AF_IUCV addressing” on page 269</td>
<td>hsuid</td>
<td>1 to 8 characters</td>
</tr>
<tr>
<td>“Setting up a HiperSockets network traffic analyzer” on page 266</td>
<td>sniffer</td>
<td>0 or 1</td>
</tr>
</tbody>
</table>

¹ not valid for HiperSockets

Tip: Use the qethconf command instead of using the attributes for IPA, proxy ARP, and VIPA directly (see “qethconf - Configure qeth devices” on page 610).

sysfs provides multiple paths through which you can access the qeth group device attributes. For example, if a device with bus ID 0.0.a100 corresponds to interface enccw0.0.a100:
all lead to the attributes for the same device. For example, the following commands are all equivalent and return the same value:

```bash
# cat /sys/bus/ccwgroup/drivers/qeth/0.0.a100/if_name
enccw0.0.a100
# cat /sys/bus/ccwgroup/devices/0.0.a100/if_name
enccw0.0.a100
# cat /sys/devices/qeth/0.0.a100/if_name
enccw0.0.a100
# cat /sys/class/net/enccw0.0.a100/device/if_name
enccw0.0.a100
```

However, the path through /sys/class/net is available only while the device is online.

**Tip:** Work through one of the paths that are based on the device bus-ID.

The following sections describe the tasks in detail.

### Creating a qeth group device

Use the `znetconf` command to configure network devices. Alternatively, you can use sysfs.

**Before you begin**

You must know the device bus-IDs that correspond to the read, write, and data subchannel of your OSA-Express CHPID in QDIO mode or HiperSockets CHPID as defined in the IOCDS of your mainframe.

**Procedure**

To create a qeth group device, either:

- Issue the `znetconf` command to create and configure a group device. The command groups the correct bus-IDs for you and sets the device online. For information about the `znetconf` command, see “znetconf - List and configure network devices” on page 648.

- Write the device numbers of the subchannel triplet to the sysfs `group` attribute to only define a group device. Issue a command of the form:

```bash
# echo <read_device_bus_id>,<write_device_bus_id>,<data_device_bus_id> > /sys/bus/ccwgroup/drivers/qeth/group
```

**Results**

The qeth device driver uses the device bus-ID of the read subchannel to create a directory for a group device:

```
/sys/bus/ccwgroup/drivers/qeth/<read_device_bus_id>
```

This directory contains a number of attributes that determine the settings of the qeth group device. The following sections describe how to use these attributes to configure a qeth group device.
Example

In this example (see Figure 51), a single OSA-Express CHPID in QDIO mode is used to connect a Linux instance to a network.

Mainframe configuration:

![Diagram of mainframe configuration](image)

**Figure 51. Mainframe configuration**

Linux configuration:

Assuming that 0.0.aa00 is the device bus-ID that corresponds to the read subchannel:

```bash
# echo 0.0.aa00,0.0.aa01,0.0.aa02 > /sys/bus/ccwgroup/drivers/qeth/group
```

This command results in the creation of the following directories in sysfs:

- `/sys/bus/ccwgroup/drivers/qeth/0.0.aa00`
- `/sys/bus/ccwgroup/devices/0.0.aa00`
- `/sys/devices/qeth/0.0.aa00`

Both the command and the resulting directories would be the same for a HiperSockets CHPID.

**Removing a qeth group device**

Use the `ungroup` sysfs attribute to remove a qeth group device.

**Before you begin**

The device must be set offline before you can remove it.

**Procedure**

To remove a qeth group device, write 1 to the `ungroup` attribute. Issue a command of the form:

```bash
echo 1 > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/ungroup
```
Example

This command removes device 0.0.aa00:

```
  echo 1 > /sys/bus/ccwgroup/drivers/qeth/0.0.aa00/ungroup
```

Setting the layer2 attribute

If the detected hardware is known to be exclusively run in a discipline, the corresponding discipline module is automatically requested.

Before you begin

- To change a configured layer2 attribute, the network interface must be shut down and the device must be set offline.
- If you are using the layer2 option within a QDIO-based guest LAN environment, you cannot define a VLAN with ID 1, because ID 1 is reserved for z/VM use.

About this task

The qeth device driver attempts to load the layer 3 discipline for HiperSockets devices and layer 2 for non-HiperSockets devices.

You can use the layer 2 mode for almost all device types, however, note the following about layer 2 to layer 3 conversion:

real OSA-Express

Hardware is able to convert layer 2 to layer 3 traffic and vice versa and thus there are no restrictions.

HiperSockets

There is no support for layer 2 to layer 3 conversion and, thus, no communication is possible between HiperSockets layer 2 interfaces and HiperSockets layer 3 interfaces. Do not include HiperSockets layer 2 interfaces and HiperSockets layer 3 interfaces in the same LAN.

z/VM guest LAN

Linux must configure the same mode as the underlying z/VM virtual LAN definition. The z/VM definition "Ethernet mode" is available for VSWITCHes and for guest LANs of type QDIO.

Procedure

The qeth device driver separates the configuration options in sysfs according to the device discipline. Hence the first configuration action after you group the device must be the configuration of the discipline. To set the discipline, issue a command of the form:

```
echo <integer> > /sys/devices/qeth/<device_bus_id>/layer2
```

where <integer> is

- 0 to turn off the layer2 attribute; this results in the layer 3 discipline.
- 1 to turn on the layer2 attribute; this results in the layer 2 discipline (default).

If the layer2 attribute has a value of -1, the layer was not set. The default layer setting is used when the device is set online.
Results

If you configured the discipline successfully, more configuration attributes are shown (for example route4 for the layer 3 discipline) and can be configured. If an OSA device is not configured for a discipline but is set online, the device driver assumes that it is a layer 2 device. It then tries to load the layer 2 discipline.

For more information about layer2, see:
- Open Systems Adapter-Express Customer’s Guide and Reference, SA22-7935
- OSA-Express Implementation Guide, SC24-5948
- Networking Overview for Linux on zSeries, REDP-3901
- z/VM Connectivity, SC24-6174

Using priority queueing

An OSA-Express CHPID in QDIO mode has up to four output queues (queues 0 - 3). The priority queueing feature gives these queues different priorities (queue 0 having the highest priority). The four output queues are available only if multiple priority is enabled for queues on the OSA-Express CHPID in QDIO mode.

Before you begin

- Priority queueing applies to OSA-Express CHPIDs in QDIO mode only.
- If more than 160 TCP/IP stacks per OSA-Express CHPID are defined in the IOCDS, priority queueing is disabled.
- The device must be offline while you set the queueing options.

About this task

Queueing is relevant mainly to high-traffic situations. When there is little traffic, queueing has no impact on processing. The qeth device driver can put data on one or more of the queues. By default, the driver uses queue 2 for all data.

Procedure

You can determine how outgoing IP packages are assigned to queues by setting a value for the priority_queueing attribute of your qeth device.

Issue a command of the form:

```
# echo <method> > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/priority_queueing
```

where `<method>` can be any of these values:

**prio_queueing_vlan**

to base the queue assignment on the two most significant bits in the priority code point in the IEEE 802.1Q header as used in VLANs. This value affects only traffic with VLAN headers, and hence works only with qeth devices in layer 2 mode.

You can set the priority code point in the IEEE 802.1Q headers of the traffic based on skb->priority by using a command of the form:

```
ip link add link <link> name <name> type vlan id <vlan-id> egress-qos-map <mapping>
```

**Note:** Enabling this option makes all traffic default to queue 3.
**prio_queueing_skb**

to base the queue assignment on the priority flag of the skbs. An skb, or socket buffer, is a Linux kernel-internal structure that represents network data. The mapping to the priority queues is as follows:

<table>
<thead>
<tr>
<th>Priority flag of the skb</th>
<th>Priority queue</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>3</td>
</tr>
<tr>
<td>2-3</td>
<td>2</td>
</tr>
<tr>
<td>4-5</td>
<td>1</td>
</tr>
<tr>
<td>≥6</td>
<td>0</td>
</tr>
</tbody>
</table>

You can use `prio_queueing_skb` for any network setups, including conventional LANs.

Use either `sockopt SO_PRIORITY` or an appropriate `iptables` command to adjust the priority flag of the skb (`skb->priority`).

Note: The priority flag of the skbs defaults to 0, hence enabling this option makes all traffic default to queue 3.

**prio_queueing_prec**

to base the queue assignment on the two most significant bits of each packet's IP header precedence field.

**prio_queueing_tos**

Deprecated; do not use for new setups.

**no_prio_queueing**

causes the qeth device driver to use queue 2 for all packets. This value is the default.

**no_prio_queueing:0**

causes the qeth device driver to use queue 0 for all packets.

**no_prio_queueing:1**

causes the qeth device driver to use queue 1 for all packets.

**no_prio_queueing:2**

causes the qeth device driver to use queue 2 for all packets. This value is equivalent to the default.

**no_prio_queueing:3**

causes the qeth device driver to use queue 3 for all packets.

**Example**

To read the current value of priority queueing for device 0.0.a110, issue:

```
# cat /sys/bus/ccwgroup/drivers/qeth/0.0.a110/priority_queueing
```

Possible results are:

**by VLAN headers**

if `prio_queueing_vlan` is set.

**by skb-priority**

if `prio_queueing_skb` is set.

**by precedence**

if `prio_queueing_prec` is set.
by type of service
  if prio_queueing_tos is set.
always queue <x>
  otherwise.

To configure queueing by skb->priority setting for device 0.0.a110 issue:

```
# echo prio_queueing_skb > /sys/bus/ccwgroup/drivers/qeth/0.0.a110/priority_queueing
```

### Specifying the number of inbound buffers

Depending on the amount of available storage and the amount of traffic, you can assign 8 - 128 inbound buffers for each qeth group device.

**Before you begin**

The device must be offline while you specify the number of inbound buffers.

**About this task**

By default, the qeth device driver assigns 64 inbound buffers to OSA devices and 128 to HiperSockets devices.

The Linux memory usage for inbound data buffers for the devices is \((\text{number of buffers}) \times (\text{buffer size})\).

The buffer size is equivalent to the frame size, which depends on the type of CHPID:

- For an OSA-Express CHPID in QDIO mode: 64 KB
- For HiperSockets: depending on the HiperSockets CHPID definition, 16 KB, 24 KB, 40 KB, or 64 KB

**Procedure**

Set the buffer_count attribute to the number of inbound buffers you want to assign. Issue a command of the form:

```
# echo <number> > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/buffer_count
```

**Example**

In this example, 64 inbound buffers are assigned to device 0.0.a000.

```
# echo 64 > /sys/bus/ccwgroup/drivers/qeth/0.0.a000/buffer_count
```

### Specifying the relative port number

Use the portno sysfs attribute to specify the relative port number.

**Before you begin**

- This description applies to adapters that, per CHPID, show more than one port to Linux.
- The device must be offline while you specify the relative port number.
Procedure

By default, the qeth group device uses port 0. To use a different port, issue a command of the form:

```
# echo <integer> > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/portno
```

Where `<integer>` is either 0 or 1.

Example

In this example, port 1 is assigned to the qeth group device.

```
# echo 1 > /sys/bus/ccwgroup/drivers/qeth/0.0.0000/portno
```

Finding out the type of your network adapter

Use the `card_type` attribute to find out the type of the network adapter through which your device is connected.

Procedure

You can find out the type of the network adapter through which your device is connected. To find out the type, read the device's `card_type` attribute. Issue a command of the form:

```
# cat /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/card_type
```

The `card_type` attribute gives information about both the type of network adapter and the type of network link (if applicable) available at the card's ports. See Table 43 for details.

<table>
<thead>
<tr>
<th>Value of card_type</th>
<th>Adapter type</th>
<th>Link type</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSD_10GIG</td>
<td>OSA card in OSD mode</td>
<td>10 Gigabit Ethernet</td>
</tr>
<tr>
<td>OSD_1000</td>
<td>OSA-Express for Unified Resource Manager</td>
<td>Gigabit Ethernet, 1000BASE-T</td>
</tr>
<tr>
<td>OSD_GbE_LANE</td>
<td></td>
<td>Gigabit Ethernet, LAN Emulation</td>
</tr>
<tr>
<td>OSD_FE_LANE</td>
<td></td>
<td>Unknown</td>
</tr>
<tr>
<td>OSD_Express</td>
<td>OSA for NCP</td>
<td>ESCON/CDLC bridge or N/A</td>
</tr>
<tr>
<td>OSM</td>
<td>OSA-Express for Unified Resource Manager</td>
<td>1000BASE-T</td>
</tr>
<tr>
<td>OSX</td>
<td>OSA-Express for zBX</td>
<td>10 Gigabit Ethernet</td>
</tr>
<tr>
<td>HiperSockets</td>
<td>HiperSockets, CHPID type IQD</td>
<td>N/A</td>
</tr>
<tr>
<td>Virtual NIC QDIO</td>
<td>VSWITCH or guest LAN based on OSA</td>
<td>N/A</td>
</tr>
<tr>
<td>Virtual NIC Hiper</td>
<td>Guest LAN based on HiperSockets</td>
<td>N/A</td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td>Other</td>
</tr>
</tbody>
</table>
Example

To find the card_type of a device 0.0.a100 issue:

```
# cat /sys/bus/ccwgroup/drivers/qeth/0.0.a100/card_type
OSD_1000
```

Setting a device online or offline

Use the online device group attribute to set a device online or offline.

Procedure

To set a qeth group device online, set the online device group attribute to 1. To set a qeth group device offline, set the online device group attribute to 0. Issue a command of the form:

```
# echo <flag> > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/online
```

Example

To set a qeth device with bus ID 0.0.a100 online issue:

```
# echo 1 > /sys/bus/ccwgroup/drivers/qeth/0.0.a100/online
```

To set the same device offline issue:

```
# echo 0 > /sys/bus/ccwgroup/drivers/qeth/0.0.a100/online
```

Finding out the interface name of a qeth group device

When a qeth group device is set online, an interface name is assigned to it.

Procedure

To find the interface name of a qeth group device, either:

- Obtain a mapping for all qeth interfaces and devices by issuing the `lsqeth -p` command.
- Find out the interface name of a qeth group device for which you know the device bus-ID by reading the group device's if_name attribute. Issue a command of the form:

```
# cat /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/if_name
```
Example

```bash
# cat /sys/bus/ccwgroup/drivers/qeth/0.0.a100/if_name
enccw0.0.a100
```

**Finding out the bus ID of a qeth interface**

Use the `lsqeth -p` command to obtain a mapping for all qeth interfaces and devices. Alternatively, you can use sysfs.

**Procedure**

To find the device bus-ID that corresponds to an interface, either:
- Use the `lsqeth -p` command.
- Use the `readlink` command. For each network interface, there is a directory in sysfs under `/sys/class/net/`, for example, `/sys/class/net/enccw0.0.f500` for interface `enccw0.0.f500`. This directory contains a symbolic link “device” to the corresponding device in `/sys/devices`. Read this link to find the device bus-ID of the device that corresponds to the interface.

Example

To find out which device bus-ID corresponds to an interface `enccw0.0.a100` issue, for example:

```
# readlink /sys/class/net/enccw0.0.a100/device
../../../0.0.a100
```

In this example, `enccw0.0.a100` corresponds to the device bus-ID `0.0.a100`.

**Activating an interface**

Use the `ip` command or equivalent to activate an interface.

**Before you begin**

- You must know the interface name of the qeth group device (see “Finding out the interface name of a qeth group device” on page 248).
- You must know the IP address that you want to assign to the device.

**About this task**

The MTU size defaults to the correct settings for HiperSockets devices. For OSA-Express CHPIDs in QDIO mode, the default MTU size depends on the device mode, layer 2 or layer 3.
- For layer 2, the default MTU is 1500 bytes.
- For layer 3, the default MTU is 1492 bytes.

In most cases, the default MTU sizes are well suited for OSA-Express CHPIDs in QDIO mode. If your network is laid out for jumbo frames, increase the MTU size to a maximum of 9000 bytes for layer 2, or to 8992 bytes for layer 3. Exceeding the defaults for regular frames or the maximum frame sizes for jumbo frames might cause performance degradation. See *Open Systems Adapter-Express Customer’s Guide and Reference*, SA22-7935 for more details about MTU size.
For HiperSockets, the maximum MTU size is restricted by the maximum frame size as announced by the Licensed Internal Code (LIC). The maximum MTU is equal to the frame size minus 8 KB. Hence, the possible frame sizes of 16 KB, 24 KB, 40 KB, or 64 KB result in maximum corresponding MTU sizes of 8 KB, 16 KB, 32 KB, or 56 KB.

The MTU size defaults to the correct settings for both HiperSockets and OSA-Express CHPIDs in QDIO mode. As a result, you do not need to specify the MTU size when you activate the interface.

On heavily loaded systems, MTU sizes that exceed 8 KB can lead to memory allocation failures for packets due to memory fragmentation. A symptom of this problem are messages of the form "order-N allocation failed" in the system log. In addition, network connections drop packets, possibly so frequently as to make the network interface unusable.

As a workaround, use MTU sizes at most of 8 KB (minus header size), even if the network hardware allows larger sizes. For example, HiperSockets or 10 Gigabit Ethernet allow larger sizes.

**Procedure**

You activate or deactivate network devices with `ip` or an equivalent command. For details of the `ip` command, see the `ip` man page.

**Examples**

- This example activates a HiperSockets CHPID with broadcast address 192.168.100.255:

  ```
  # ip addr add 192.168.100.10/24 dev enccw0.0.a1c0
  # ip link set dev enccw0.0.a1c0 up
  ```

- This example activates an OSA-Express CHPID in QDIO mode with broadcast address 192.168.100.255:

  ```
  # ip addr add 192.168.100.11/24 dev enccw0.0.f500
  # ip link set dev enccw0.0.f500 up
  ```

- This example reactivates an interface that was already activated and subsequently deactivated:

  ```
  # ip link set dev enccw0.0.f500 up
  ```

**Confirming that an IP address has been set under layer 3**

There may be circumstances that prevent an IP address from being set, most commonly if another system in the network has set that IP address already.

**About this task**

The Linux network stack design does not allow feedback about IP address changes. If `ip` or an equivalent command fails to set an IP address on an OSA-Express network CHPID, a query with `ip` shows the address as being set on the interface although the address is not actually set on the CHPID.
There are usually failure messages about not being able to set the IP address or duplicate IP addresses in the kernel messages. You can find these messages in the output of the `dmesg` command. In Red Hat Enterprise Linux 7.5, you can also find the messages in `/var/log/messages`.

If you are not sure whether an IP address was set properly or experience a networking problem, check the messages or logs to see if an error was encountered when setting the address. This also applies in the context of HiperSockets and to both IPv4 and IPv6 addresses. It also applies to whether an IP address has been set for IP takeover, for VIPA, or for proxy ARP.

**Duplicate IP addresses**
The OSA-Express adapter in QDIO mode recognizes duplicate IP addresses on the same OSA-Express adapter or in the network using ARP and prevents duplicates.

**About this task**
Several setups require duplicate addresses:

- To perform IP takeover you need to be able to set the IP address to be taken over. This address exists prior to the takeover. See “Taking over IP addresses” on page 263 for details.
- For proxy ARP you need to register an IP address for ARP that belongs to another Linux instance. See “Configuring a device for proxy ARP” on page 267 for details.
- For VIPA you need to assign the same virtual IP address to multiple devices. See “Configuring a device for virtual IP address (VIPA)” on page 268 for details.

You can use the `qethconf` command (see “qethconf - Configure qeth devices” on page 610) to maintain a list of IP addresses that your device can take over, a list of IP addresses for which your device can handle ARP, and a list of IP addresses that can be used as virtual IP addresses, regardless of any duplicates on the same OSA-Express adapter or in the LAN.

**Deactivating an interface**
You can deactivate an interface with `ip` or an equivalent command or by setting the network device offline.

**About this task**
Setting a device offline involves actions on the attached device, but deactivating a device only stops the interface logically within Linux.

**Procedure**
To deactivate an interface with `ip`, issue a command of the form:

```
# ip link set dev <interface_name> down
```

**Example**
To deactivate enccw0.0.f500 issue:

```
# ip link set dev enccw0.0.f500 down
```
Recovering a device

You can use the recover attribute of a qeth group device to recover it in case of failure.

About this task

For example, error messages in /var/log/messages might inform you of a malfunctioning device.

Procedure

Issue a command of the form:

```bash
# echo 1 > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/recover
```

Example

```bash
# echo 1 > /sys/bus/ccwgroup/drivers/qeth/0.0.a100/recover
```

Configuring checksum offload operations

Some operations can be offloaded to the OSA adapter, thus relieving the burden on the host CPU. The qeth device driver supports checksum offloading for IP, TCP and UDP network packets.

The qeth device driver supports offloading the following checksum operations on both layer 2 and layer 3:
- Inbound (receive) checksum calculations
- Outbound (send) checksum calculations

The qeth device driver also supports offloading TSO segmentation, see "Enabling and disabling TCP segmentation offload" on page 262.

VLAN interfaces inherit offload settings from their base interface.

The offload operations can be set with the Linux `ethtool` command. See the `ethtool` man page for details. The following abbreviated example shows some offload settings:
Turning inbound checksum calculations on and off

A checksum calculation is a form of redundancy check to protect the integrity of data. In general, checksum calculations are used for network data.

Procedure

The qeth device driver supports offloading checksum calculations on inbound packets to the OSA feature. To enable or disable checksum calculations by the OSA feature, issue a command of this form:

```
# ethtool -K <interface_name> rx <value>
```

where `<value>` is on or off.

Examples

- To let the OSA feature calculate the inbound checksum for network device enccw0.0.f500, issue

  ```
  # ethtool -K enccw0.0.f500 rx on
  ```

- To let the host CPU calculate the inbound checksum for network device enccw0.0.f500, issue

  ```
  # ethtool -K enccw0.0.f500 rx off
  ```

Turning outbound checksum calculations on and off

The qeth device driver supports offloading outbound (send) checksum calculations to the OSA feature.

About this task

You can enable or disable the OSA feature that calculates the outbound checksums by using the `ethtool` command.
Attention:  For OSA-Express3 and earlier: When outbound checksum calculations are offloaded, the OSA feature performs the checksum calculations. Offloaded checksum calculations only applies to packets that go out to the LAN or come in from the LAN. Linux instances that share an OSA port exchange packages directly. The packages are forwarded by the OSA adapter but do not go out on the LAN and no checksum offload is performed. The qeth device driver cannot detect this, and so cannot issue any warning about it.

Procedure

Issue a command of the form:

```
# ethtool -K <interface_name> tx <value>
```

where `<value>` is on or off.

Examples

- To let the OSA feature calculate the outbound checksum for network device enccw0.0.f500, issue

  ```
  # ethtool -K enccw0.0.f500 tx on
  ```

- To let the host CPU calculate the outbound checksum for network device enccw0.0.f500, issue

  ```
  # ethtool -K enccw0.0.f500 tx off
  ```

Isolating data connections

You can restrict communications between operating system instances that share an OSA port on an OSA adapter.

About this task

A Linux instance can configure the OSA adapter to prevent any direct package exchange between itself and other operating system instances that share an OSA adapter. This configuration ensures a higher degree of isolation than VLANs.

QDIO data connection isolation is configured as a policy. The policy is implemented as a sysfs attribute called `isolation`. The attribute appears in `sysfs` regardless of whether the hardware supports the feature. The policy can take the following values:

- **none**  No isolation. This is the default.
- **drop**  Specifies the ISOLATION_DROP policy. All packets from guests sharing the same OSA adapter to the guest having this policy configured are dropped automatically. The same holds for all packets sent by the guest having this policy configured to guests on the same OSA card. All packets to or from the isolated guest need to have a target that is not hosted on the OSA card. You can accomplish this by a router hosted on a separate machine or a separate OSA adapter.

  For example, assume that three Linux instances share an OSA adapter, but only one instance (Linux A) needs to be isolated. Then Linux A declares its OSA adapter (QDIO Data Connection to the OSA adapter) to be isolated. Any packet being sent to or from Linux A must pass at least the physical
switch to which the shared OSA adapter is connected. Linux A cannot communicate with other instances that share the OSA adapter, here B or C. The two other instances could still communicate directly through the OSA adapter without the external switch in the network path (see Figure 52).

Figure 52. Linux instance A is isolated from instances B and C

**forward**

Specifies the ISOLATION_FORWARD policy. All packets are passed through a switch. The ISOLATION_FORWARD policy requires a network adapter in Virtual Ethernet Port Aggregator (VEPA) mode with an adjacent switch port configured for reflective relay mode.

To check whether the switch of the adapter is in reflective relay mode, read the sysfs attribute switch_attrs. The attribute lists all supported forwarding modes, with the currently active mode enclosed in square brackets. For example:

```
cat /sys/devices/qeth/0.0.f5f0/switch_attrs
802.1 [rr]
```

The example indicates that the adapter supports both 802.1 forwarding mode and reflective relay mode, and reflective relay mode is active.

Using a network adapter in VEPA mode achieves further isolation. VEPA mode forces traffic from the Linux guests to be handled by the external switch. For example, Figure 53 on page 256 shows instances A and B with ISOLATION_FORWARD specified for the policy. All traffic between A and B goes through the external switch. The rule set of the switch now determines which connections are possible. The graphic assumes that A can communicate with B, but not with C.
If the ISOLA TION_FOR WARD policy was enforced successfully, but the switch port later loses the reflective-relay capability, the device is set offline to prevent damage.

You can configure the policy regardless of whether the device is online. If the device is online, the policy is configured immediately. If the device is offline, the policy is configured when the device comes online.

**Examples**

- To check the current isolation policy:

  ```
  # cat /sys/devices/qeth/0.0.f5f0/isolation
  ```

- To set the isolation policy to ISOLA TION_DROP:

  ```
  # echo "drop" > /sys/devices/qeth/0.0.f5f0/isolation
  ```

- To set the isolation policy to ISOLA TION_FORWARD:

  ```
  # echo "forward" > /sys/devices/qeth/0.0.f5f0/isolation
  ```

If the switch is not capable of VE PA support, or VE PA support is not configured on the switch, then you cannot set the isolation attribute value to 'forward' while the device is online. If the switch does not support VE PA and you set the isolation value 'forward' while the device is offline, then the device cannot be set online until the isolation value is set back to 'drop' or 'none'.

- To set the isolation policy to none:

  ```
  # echo "none" > /sys/devices/qeth/0.0.f5f0/isolation
  ```

When using vNICs, VE PA mode needs to be enabled on the respective VSWITCH. See z/VM Connectivity, SC24-6174 for information about setting up data connection isolation on a VSWITCH.
Starting and stopping collection of QETH performance statistics

Use the performance_stats attribute to start and stop collection of QETH performance statistics.

About this task

For QETH performance statistics, there is a device group attribute called

This attribute is initially set to 0, that is, QETH performance data is not collected.

Procedure

To start collection for a specific QETH device, write 1 to the attribute. For example:

```
echo 1 > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/performance_stats
```

To stop collection write 0 to the attribute, for example:

```
echo 0 > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/performance_stats
```

Stopping QETH performance data collection for a specific QETH device is accompanied by a reset of current statistic values to zero.

To display QETH performance statistics, use the `ethtool` command. See the `ethtool` man page for details.
Example

The following example shows statistic and device driver information:

```bash
# ethtool -S enccw0.0.f500
NIC statistics:
  rx skbs: 86
  rx buffers: 85
  tx skbs: 86
  tx buffers: 86
  tx skb no packing: 86
  tx buffers no packing: 86
  tx skb packing: 0
  tx buffers packing: 0
  tx sg skbs: 0
  tx sg frags: 0
  rx skbs: 0
  rx sg skbs: 0
  rx sg frags: 0
  rx sg page allocs: 0
  tx large kbytes: 0
  tx large count: 0
  tx pk state ch n->p: 0
  tx pk state ch p->n: 0
  tx pk watermark low: 2
  tx pk watermark high: 5
  queue 0 buffer usage: 0
  queue 1 buffer usage: 0
  queue 2 buffer usage: 0
  queue 3 buffer usage: 0
  rx handler time: 856
  rx handler count: 84
  rx do_QDIO time: 16
  rx do_QDIO count: 11
  tx handler time: 330
  tx handler count: 87
  tx time: 1236
  tx count: 86
  tx do_QDIO time: 997
  tx do_QDIO count: 86
  tx csum: 0
  tx lin: 0
  cq handler count: 0
  cq handler time: 0

# ethtool -i enccw0.0.f500
driver: qeth_13
version: 1.0
firmware-version: 087a
bus-info: 0.0.f5f0/0.0.f5f1/0.0.f5f2
supports-statistics: yes
supports-test: no
supports-eprom-access: no
supports-register-dump: no
supports-priv-flags: no
```

Capturing a hardware trace

Hardware traces are intended for use by the IBM service organization. Hardware tracing is turned off by default. Turn on the hardware-tracing feature only when instructed to do so by IBM service.

Before you begin

- The OSA-Express adapter must support the hardware-tracing feature.
- The qeth device must be online to return valid values of the hw_trap attribute.
About this task

When errors occur on an OSA-Express adapter, both software and hardware traces must be collected. The hardware-tracing feature requests a hardware trace if an error is detected. This feature makes it possible to correlate the hardware trace with the device driver trace. If the hardware-tracing feature is activated, traces are captured automatically, but you can also start the capturing yourself.

Procedure

To activate or deactivate the hardware-tracing feature, issue a command of the form:

```
# echo <value> > /sys/devices/qeth/<device_bus_id>/hw_trap
```

Where `<value>` can be:

- **arm**: If the hardware-tracing feature is supported, write arm to the `hw_trap` sysfs attribute to activate it. If the hardware-tracing feature is present and activated, the `hw_trap` sysfs attribute has the value `arm`.

- **disarm**: Write disarm to the `hw_trap` sysfs attribute to turn off the hardware-tracing feature. If the hardware-tracing feature is not present or is turned off, the `hw_trap` sysfs attribute has the value `disarm`. This setting is the default.

- **trap**: (Write only) Capture a hardware trace. Hardware traces are captured automatically, but if asked to do so by IBM service, you can start the capturing yourself by writing `trap` to the `hw_trap` sysfs attribute. The hardware trap function must be set to arm.

Examples

In this example the hardware-tracing feature is activated for qeth device 0.0.a000:

```
# echo arm > /sys/devices/qeth/0.0.a000/hw_trap
```

In this example a trace capture is started on qeth device 0.0.a000:

1. Check that the `hw_trap` sysfs attribute is set to arm:

   ```
   # cat /sys/devices/qeth/0.0.a000/hw_trap
   arm
   ```

2. Start the capture:

   ```
   # echo trap > /sys/devices/qeth/0.0.a000/hw_trap
   ```

Working with qeth devices in layer 3 mode

Tasks you can perform on qeth devices in layer 3 mode include setting up a router, configuring offload operations, and taking over IP addresses.

Use the layer 2 attribute to set the mode. See “Setting the layer2 attribute” on page 243 about setting the mode. See “Layer 2 and layer 3” on page 229 for general information about the layer 2 and layer 3 disciplines.
Setting up a Linux router in layer 3

By default, your Linux instance is not a router. Depending on your IP version, IPv4 or IPv6 you can use the route4 or route6 attribute of your qeth device to define it as a router.

Before you begin

- A suitable hardware setup must be in place that enables your Linux instance to act as a router.
- The Linux instance is set up as a router. To configure Linux running in a z/VM guest virtual machine or in an LPAR as a router, IP forwarding must be enabled in addition to setting the route4 or route6 attribute.

For IPv4, enable IP forwarding by issuing:

```
# sysctl -w net.ipv4.conf.all.forwarding=1
```

For IPv6, enable IP forwarding by issuing:

```
# sysctl -w net.ipv6.conf.all.forwarding=1
```

About this task

You can set the route4 or route6 attribute dynamically, while the qeth device is online.

The same values are possible for route4 and route6 but depend on the type of CHPID:

Table 44. Summary of router setup values

<table>
<thead>
<tr>
<th>Router specification</th>
<th>OSA-Express CHPID in QDIO mode</th>
<th>HiperSockets CHPID</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary_router</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>secondary_router</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>primary_connector</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>secondary_connector</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>multicast_router</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>no_router</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Both types of CHPIDs accept:

multicast_router causes the qeth driver to receive all multicast packets of the CHPID. For a unicast function for HiperSockets see “HiperSockets Network Concentrator” on page 280.

no_router is the default. You can use this value to reset a router setting to the default.

An OSA-Express CHPID in QDIO mode accepts the following values:

primary_router to make your Linux instance the principal connection between two networks.
**secondary_router**

to make your Linux instance a backup connection between two networks.

A HiperSockets CHPID accepts the following values, if the microcode level supports the feature:

**primary_connector**

to make your Linux instance the principal connection between a HiperSockets network and an external network (see “HiperSockets Network Concentrator” on page 280).

**secondary_connector**

to make your Linux instance a backup connection between a HiperSockets network and an external network (see “HiperSockets Network Concentrator” on page 280).

**Example**

In this example, two Linux instances, “Linux P” and “Linux S”, running on an IBM mainframe use OSA-Express to act as primary and secondary routers between two networks. IP forwarding must be enabled for Linux in an LPAR or as a z/VM guest to act as a router. In Red Hat Enterprise Linux 7.5 you can set IP forwarding permanently in /etc/sysctl.conf or dynamically with the `sysctl` command.

**Mainframe configuration:**

![Figure 54. Mainframe configuration](image)

It is assumed that both Linux instances are configured as routers in their LPARs or in z/VM.

**Linux P configuration:**

To create the qeth group devices:

```
# echo 0.0.0400,0.0.0401,0.0.0402 > /sys/bus/ccwgroup/drivers/qeth/group
# echo 0.0.0200,0.0.0201,0.0.0202 > /sys/bus/ccwgroup/drivers/qeth/group
```

To make Linux P a primary router for IPv4:

```
# echo primary_router > /sys/bus/ccwgroup/drivers/qeth/0.0.0400/route4
# echo primary_router > /sys/bus/ccwgroup/drivers/qeth/0.0.0200/route4
```

**Linux S configuration:**

To create the qeth group devices:
To make Linux S a secondary router for IPv4:

```
# echo secondary_router > /sys/bus/ccwgroup/drivers/qeth/0.0.0404/route4
# echo secondary_router > /sys/bus/ccwgroup/drivers/qeth/0.0.0204/route4
```

In this example, qeth device 0.01510 is defined as a primary router for IPv6:

```
# echo 0.0.0404,0.0.0405,0.0.0406 > /sys/bus/ccwgroup/drivers/qeth/group
# echo 0.0.0204,0.0.0205,0.0.0206 > /sys/bus/ccwgroup/drivers/qeth/group
```

See “HiperSockets Network Concentrator” on page 280 for further examples.

**Enabling and disabling TCP segmentation offload**

Offloading the TCP segmentation operation from the Linux network stack to the adapter can lead to enhanced performance for interfaces with predominately large outgoing packets.

**Procedure**

To support TCP segmentation offload (TSO), a network device must support outbound (TX) checksumming and scatter gather. For this reason, you must turn on scatter gather and outbound checksumming prior to configuring TSO. All three options can be turned on or off with a single `ethtool` command of the form:

```
# ethtool -k <interface_name> tx <value> sg <value> tso <value>
```

where `<value>` is either on or off.

**Note:** When TCP segmentation is offloaded, the OSA feature performs the calculations. Offloaded calculations apply only to packets that go out to the LAN or come in from the LAN. Linux instances that share an OSA port exchange packages directly. The packages are forwarded by the OSA adapter but do not go out on the LAN and no TCP segmentation calculation is performed. The qeth device driver cannot detect this, and so cannot issue any warning about it.

**Examples**

- To offload the TCP segmentation operation for a network device enccw0.0.f500 issue:

  ```
  # ethtool -K enccw0.0.f500 tx on sg on tso on
  ```

- To disable offloading the TCP segmentation operation for a network device enccw0.0.f500 issue:

  ```
  # ethtool -K enccw0.0.f500 tx off sg off tso off
  ```
Faking broadcast capability

It is possible to fake the broadcast capability for devices that do not support broadcasting.

Before you begin

- You can fake the broadcast capability only on devices that do not support broadcast.
- The device must be offline while you enable faking broadcasts.

About this task

For devices that support broadcast, the broadcast capability is enabled automatically.

To find out whether a device supports broadcasting, use the `ip` command. If the resulting list shows the BROADCAST flag, the device supports broadcast. This example shows that the device enccw0.0.f500 supports broadcast:

```
# ip -s link show dev enccw0.0.f500
3: enccw0.0.f500: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1492 qdisc pfifo_fast qlen 1000
   link/ether 00:11:25:bd:da:66 brd ff:ff:ff:ff:ff:ff
RX: bytes packets errors dropped overrun mcast
   236350  2974   0   0   0    9
TX: bytes packets errors dropped carrier collsns
   374443  1791   0   0   0    0
```

Some processes, for example, the `gated` routing daemon, require the devices' broadcast capable flag to be set in the Linux network stack.

Procedure

To set the broadcast capable flag for devices that do not support broadcast, set the fake_broadcast attribute of the qeth group device to 1. To reset the flag, set it to 0. Issue a command of the form:

```
# echo <flag> > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/fake_broadcast
```

Example

In this example, a device 0.0.a100 is instructed to pretend that it can broadcast.

```
# echo 1 > /sys/bus/ccwgroup/drivers/qeth/0.0.a100/fake_broadcast
```

Taking over IP addresses

You can configure IP takeover if the layer2 option is not enabled. If you enabled the layer2 option, you can configure for IP takeover as you would in a distributed server environment.

About this task

For information about the layer2 option, see "MAC headers in layer 2 mode" on page 231.
Taking over an IP address overrides any previous allocation of this address to another LPAR. If another LPAR on the same CHID already registered for that IP address, this association is removed.

An OSA-Express CHID in QDIO mode can take over IP addresses from any IBM Z operating system. IP takeover for HiperSockets CHIDs is restricted to taking over addresses from other Linux instances in the same Central Electronics Complex (CEC).

IP address takeover between multiple CHIDs requires ARP for IPv4 and Neighbor Discovery for IPv6. OSA-Express handles ARP transparently, but not Neighbor Discovery.

There are three stages to taking over an IP address:

Stage 1: Ensure that your qeth group device is enabled for IP takeover
Stage 2: Activate the address to be taken over for IP takeover
Stage 3: Issue a command to take over the address

Stage 1: Enabling a qeth group device for IP takeover
For OSA-Express and HiperSockets CHIDs, both the qeth group device that is to take over an IP address and the device that surrenders the address must be enabled for IP takeover.

Procedure

By default, qeth devices are not enabled for IP takeover. To enable a qeth group device for IP address takeover set the enable device group attribute to 1. To switch off the takeover capability set the enable device group attribute to 0. In sysfs, the enable attribute is located in a subdirectory ipa_takeover. Issue a command of the form:

```
# echo <flag> > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/ipa_takeover/enable
```

Example

In this example, a device 0.0.a500 is enabled for IP takeover:

```
# echo 1 > /sys/bus/ccwgroup/drivers/qeth/0.0.a500/ipa_takeover/enable
```

Stage 2: Activating and deactivating IP addresses for takeover
The qeth device driver maintains a list of IP addresses that qeth group devices can take over or surrender. To enable Linux to take over an IP-address or to surrender an address, the address must be added to this list.

Procedure

Use the qethconf command to add IP addresses to the list.

- To display the list of IP addresses that are activated for IP takeover issue:

  ```
  # qethconf ipa list
  ```

- To activate an IP address for IP takeover, add it to the list. Issue a command of the form:

  ```
  # qethconf ipa activate <ip_address>
  ```
To deactivate an IP address delete it from the list. Issue a command of the form:

```
# qethconf ipa del <ip_address>/<mask_bits> <interface_name>
```

In these commands, `<ip_address>/<mask_bits>` is the range of IP addresses to be activated or deactivated. See "qethconf - Configure qeth devices" on page 610 for more details about the qethconf command.

**IPv4 example:**

In this example, there is only one range of IP addresses (192.168.10.0 to 192.168.10.255) that can be taken over by HiperSockets device enccw0.0.a1c0.

List the range of IP addresses (192.168.10.0 to 192.168.10.255) that can be taken over by HiperSockets device enccw0.0.a1c0.

```
# qethconf ipa list
ipa add 192.168.10.0/24 enccw0.0.a1c0
```

The following command adds a range of IP addresses that can be taken over by OSA device enccw0.0.f500.

```
# qethconf ipa add 192.168.11.0/24 enccw0.0.f500
qethconf: Added 192.168.11.0/24 to /sys/class/net/enccw0.0.f500/device/ipa_takeover/add4.
qethconf: For verification please use "qethconf ipa list"
```

Listing the activated IP addresses now shows both ranges of addresses.

```
# qethconf ipa list
ipa add 192.168.10.0/24 enccw0.0.a1c0
ipa add 192.168.11.0/24 enccw0.0.f500
```

The following command deletes the range of IP addresses that can be taken over by OSA device enccw0.0.f500.

```
# qethconf ipa del 192.168.11.0/24 enccw0.0.f500
qethconf: Deleted 192.168.11.0/24 from
    sysfs entry /sys/class/net/enccw0.0.f500/device/ipa_takeover/del4.
qethconf: For verification please use "qethconf ipa list"
```

**IPv6 example:**

The following command adds one range of IPv6 addresses, fec0:0000:0000:0000:0000:0000:0000:0000 to fec0:0000:0000:0000:0000:0000:FFFF:FFFF, that can be taken over by device enccw0.0.d300.

Add a range of IP addresses:

```
qethconf ipa add fec0::/64 enccw0.0.d300
qethconf: Added fec0:0000:0000:0000:0000:0000:0000:0000 to
    sysfs entry /sys/class/net/enccw0.0.d300/device/ipa_takeover/add6.
qethconf: For verification please use "qethconf ipa list"
```
Listing the activated IP addresses now shows the range of addresses:

```
qethconf ipa list
...
ipa add fec0:0000:0000:0000:0000:0000:0000:0000/64 enccw0.0.d300
```

The following command deletes the IPv6 address range that can be taken over by enccw0.0.d300:

```
qethconf ipa del fec0:0000:0000:0000:0000:0000:0000:0000/64 enccw0.0.d300:
qethconf: Deleted fec0:0000:0000:0000:0000:0000:0000:0000/64 from
sysfs entry /sys/class/net/enccw0.0.d300/device/ipa_takeover/del6.
qethconf: For verification please use "qethconf ipa list"
```

**Stage 3: Issuing a command to take over the address**

To complete taking over a specific IP address and remove it from the CHPID or LPAR that previously held it, issue an `ip addr` or equivalent command.

**Before you begin**

- Both the device that is to take over the IP address and the device that is to surrender the IP address must be enabled for IP takeover. This rule applies to the devices on both OSA-Express and HiperSockets CHPIDs. (See “Stage 1: Enabling a qeth group device for IP takeover” on page 264).
- The IP address to be taken over must have been activated for IP takeover (see “Stage 2: Activating and deactivating IP addresses for takeover” on page 264).

**About this task**

Be aware of the information in “Confirming that an IP address has been set under layer 3” on page 250 when using IP takeover.

**Examples**

**IPv4 example:**

To make a HiperSockets device enccw0.0.a1c0 take over IP address 192.168.10.22 issue:

```
# ip addr add 192.168.10.22/24 dev enccw0.0.a1c0
```

For IPv4, the IP address you are taking over must be different from the one that is already set for your device. If your device already has the IP address it is to take over, you must issue two commands: First remove the address to be taken over if it is already there. Then add the IP address to be taken over.

For example, to make a HiperSockets device enccw0.0.a1c0 take over IP address 192.168.10.22 if enccw0.0.a1c0 is already configured to have IP address 192.168.10.22 issue:

```
# ip addr del 192.168.10.22/24 dev enccw0.0.a1c0
# ip addr add 192.168.10.22/24 dev enccw0.0.a1c0
```

**IPv6 example:**

To make a device enccw0.0.d300 take over fec0:111:25ff:feb:d9da/64 issue:

```
```
For IPv6, setting the `nodad` (no duplicate address detection) option ensures that the enccw0.0.d300 interface uses the IP address fec0::111:25ff:febd:d9da/64. Without the `nodad` option, the previous owner of the IP address might prevent the takeover by responding to a duplicate address detection test.

The IP address you are taking over must be different from the one that is already set for your device. If your device already has the IP address it is to take over you must issue two commands: First remove the address to be taken over if it is already there. Then add the IP address to be taken over.

For example, to make a device enccw0.0.d300 take over IP address fec0::111:25ff:febd:d9da/64 when enccw0.0.d300 is already configured to have that particular IP address issue:

```
ip addr del fec0::111:25ff:febd:d9da/64
ip addr add fec0::111:25ff:febd:d9da/64 nodad dev enccw0.0.d300
```

### Configuring a device for proxy ARP

You can configure a device for proxy ARP if the layer2 option is not enabled. If you enabled the layer2 option, you can configure for proxy ARP as you would in a distributed server environment.

**Before you begin**

Configure only qeth group devices that are set up as routers for proxy ARP.

**About this task**

For information about the layer2 option, see “MAC headers in layer 2 mode” on page 231.


Use the `qethconf` command to display this list or to change the list by adding and removing IP addresses (see “qethconf - Configure qeth devices” on page 610).

Be aware of the information in “Confirming that an IP address has been set under layer 3” on page 250 when you work with proxy ARP.

**Example**

Figure 55 on page 268 shows an environment where proxy ARP is used.
G1, G2, and G3 are instances of Linux on z/VM (connected, for example, through a guest LAN to a Linux router R), reached from GW (or the outside world) through R. R is the ARP proxy for G1, G2, and G3. That is, R agrees to take care of packets that are destined for G1, G2, and G3. The advantage of using proxy ARP is that GW does not need to know that G1, G2, and G3 are behind a router.

To receive packets for 1.2.3.4, so that it can forward them to G1 1.2.3.4, R would add 1.2.3.4 to its list of IP addresses for proxy ARP for the interface that connects it to the OSA adapter.

```
# qethconf parp add 1.2.3.4 enccw0.0.f500
qethconf: Added 1.2.3.4 to /sys/class/net/enccw0.0.f500/device/rxip/add4.
qethconf: Use "qethconf parp list" to check for the result
```

After issuing similar commands for the IP addresses 1.2.3.5 and 1.2.3.6 the proxy ARP configuration of R would be:

```
# qethconf parp list
parp add 1.2.3.4 enccw0.0.f500
parp add 1.2.3.5 enccw0.0.f500
parp add 1.2.3.6 enccw0.0.f500
```

### Configuring a device for virtual IP address (VIPA)

You can configure a device for VIPA if the layer2 option is not enabled. If you enabled the layer2 option, you can configure for VIPA as you would in a distributed server environment.

#### About this task

For information about the layer2 option, see “MAC headers in layer 2 mode” on page 231.

IBM Z uses VIPAs to protect against certain types of hardware connection failure. You can assign VIPAs that are independent from particular adapter. VIPAs can be built under Linux using `dummy` devices (for example, “dummy0” or “dummy1”).

The qeth device driver maintains a list of VIPAs that the OSA-Express adapter accepts for each qeth group device. Use the `qethconf` utility to add or remove VIPAs (see “qethconf - Configure qeth devices” on page 610).

For an example of how to use VIPA, see “Scenario: VIPA – minimize outage due to adapter failure” on page 272.
Be aware of “Confirming that an IP address has been set under layer 3” on page 250 when you work with VIPAs.

Configuring a HiperSockets device for AF_IUCV addressing

Use the hsuid attribute of a HiperSockets device in layer 3 mode to identify it to the AF_IUCV addressing family support.

Before you begin

- Support for AF_IUCV based connections through real HiperSockets requires Completion Queue Support.
- The device must be set up for AF_IUCV addressing (see “Setting up HiperSockets devices for AF_IUCV addressing” on page 320).

Procedure

To set an identifier, issue a command of this form:

```
# echo <value> > /sys/bus/ccwgroup/drivers/qeth/0.0.a007/hsuid
```

The identifier is case-sensitive and must adhere to these rules:

- It must be 1 - 8 characters.
- It must be unique across your environment.
- It must not match any z/VM user ID in your environment. The AF_IUCV addressing family support also supports z/VM IUCV connections.

Example

In this example, MYHOST01 is set as the identifier for a HiperSockets device with bus ID 0.0.a007.

```
# echo MYHOST01 > /sys/bus/ccwgroup/drivers/qeth/0.0.a007/hsuid
```

Working with qeth devices in layer 2 mode

Tasks that you can perform on qeth devices in layer 2 mode include setting up an OSA or HiperSockets bridge port and configuring notification behavior for the bridge port.

Use the layer2 attribute to set the mode. See “Setting the layer2 attribute” on page 243 about setting the mode. See “Layer 2 and layer 3” on page 229 for general information about the layer 2 and layer 3 disciplines.

Configuring a network device as a member of a Linux bridge

You can define an OSA or HiperSockets device to be a bridge port, which allows it to act as a member of a Linux software bridge. Use the bridge_role attribute of a network device in layer 2 to make it receive all traffic with unknown destination MAC addresses.

Before you begin

To use the bridging support, you need OSA or HiperSockets hardware that supports layer 2 SETBRIDGEPORT functionality.
You can have one active bridge port per Internal Queued Direct Communication (IQD) channel. You can have either only secondary bridge ports, or one primary and several secondary bridge ports.

A HiperSockets bridge port requires that Linux runs as a z/VM guest.

For more information about the bridge port concept, see “Layer 2 bridge port function” on page 235.

About this task

The following sysfs attributes control the bridge port functions. The attributes can be found in the /sys/bus/ccwgroup/drivers/qeth/<device_bus_id> directory.

**bridge_role**

Read-write attribute that controls the role of the port. Valid values are:

- **primary**
  - Assigns the port the primary bridge port role.

- **secondary**
  - Assigns the port a secondary bridge port role.

- **none**
  - Revokes existing bridge port roles and indicates that no role is assigned.

Assigning a role directly to a port prevents use of the bridge_reflect_promisc attribute.

**bridge_state**

Read-only attribute that shows the state of the port. Valid values are:

- **active**
  - The port is assigned a bridge port role and is switched into active state by the adapter. The device receives frames that are addressed to unknown MAC addresses.

- **standby**
  - The port is assigned a bridge port role, but is not currently switched into active state by the adapter. The device does not receive frames that are destined to unknown MAC addresses.

- **inactive**
  - The port is not assigned a bridge port role.

**bridge_hostnotify**

HiperSockets only: Read-write attribute that controls the sending of notifications for the port. When you enable notifications (even if notifications were already enabled), udev events are emitted for all currently connected communication peers in quick succession. After that, a udev event is emitted every time a communication peer is connected, or a previously connected peer is disconnected. Any user space program that monitors these events must repopulate its list of registered peers every time the status of the bridge port device changes to enable notifications.

Valid values are:

- **1**
  - The port is set to send notifications.

- **0**
  - Notifications are turned off.

Notifications about the change of the state of bridge ports, and (if enabled) about registration and deregistration of communication peers on the LAN are
delivered as udev events. The events are described in the file Documentation/s390/qeth.txt in the Linux kernel source tree.

**bridge_reflect_promisc**

Read-write attribute that, when set, makes the bridge-port role of the port follow ("reflect") the promiscuity flag (IFF_PROMISC) of the corresponding Linux network interface. You can specify the following values:

- **none** Setting and resetting the promiscuous mode on the network interface has no effect on the bridge-port role of the underlying port.

- **primary** Setting or resetting the promiscuous mode on the network interface that is served by this device causes the driver to attempt assigning (or resetting) the primary role to the port. If a port with the primary role exists, assignment fails.

- **secondary** Setting or resetting the promiscuous mode on the network interface that is served by this device causes the driver to attempt assigning (or resetting) the secondary role to the port.

Setting **bridge_reflect_promisc** to anything but **none** causes the **bridge_role** attribute to become read-only. The role of a port changes as a result of setting or unsetting the promiscuity flag (IFF_PROMISC) of the corresponding network interface. You can check the currently assigned role by reading the **bridge_role** attribute.

**Procedure**

1. To configure a network device as a bridge, issue a command of this form:

   ```
   # echo <value> > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/bridge_role
   ```

   Setting the **bridge_role** attribute requires the **bridge_reflect_promisc** attribute to be **none**. Alternatively, to make the bridge-port role of the port follow the promiscuity flag (IFF_PROMISC) of the corresponding Linux network interface, issue a command of the following form:

   ```
   # echo <value> > /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/bridge_reflect_promisc
   ```

   where valid values are:

   - primary
   - secondary
   - none

2. Check the state of the bridge port by reading the **bridge_state** attribute. Issue a command of this form:

   ```
   # cat /sys/bus/ccwgroup/drivers/qeth/<device_bus_id>/bridge_state
   ```

   where displayed values could be:

   - active
   - standby
   - inactive
Example

In this example, a network device with bus ID 0.0.a007 is defined as a primary bridge port.

```
# echo primary > /sys/bus/ccwgroup/drivers/qeth/0.0.a007/bridge_role
# cat /sys/bus/ccwgroup/drivers/qeth/0.0.a007/bridge_state
active
```

What to do next

You can specify up to four secondary bridge ports together with one primary bridge port. If the primary bridge port fails, one of these bridge ports takes over. For each secondary bridge port, set bridge_role to secondary.

Scenario: VIPA – minimize outage due to adapter failure

Using VIPA you can assign IP addresses that are not associated with a particular adapter. VIPA thus minimizes outage that is caused by adapter failure.

This scenario describes how to use:

- Standard VIPA
- Source VIPA (version 2.0.0 and later)

Standard VIPA is sufficient for applications, such as web servers, that do not open connections to other nodes. Source VIPA is used for applications that open connections to other nodes. Use Source VIPA Extensions to work with multiple VIPAs per destination in order to achieve multipath load balancing.

Note:

1. See the information in “Confirming that an IP address has been set under layer 3” on page 250 concerning possible failure when you set IP addresses for OSA-Express features in QDIO mode (qeth driver).
2. The configuration file layout for Source VIPA changed since the 1.x versions. In the 2.0.0 version a policy is included. For details, see the readme file and the man pages that are provided with the package.

Standard VIPA

VIPA is a facility for assigning an IP address to a system, instead of to individual adapters. It is supported by the Linux kernel. The addresses can be in IPv4 or IPv6 format.

Setting up standard VIPA

To set up VIPA you must create a dummy device, ensure that your service listens to the IP address, and set up routing to it.

Procedure

Follow these main steps to set up VIPA in Linux:

1. Create a dummy device with a virtual IP address.
2. Ensure that your service (for example, the Apache web server) listens to the virtual IP address assigned in step 1.
3. Set up routes to the virtual IP address, on clients or gateways. To do so, you can use either:
- Static routing (shown in "Example of how to set up standard VIPA").
- Dynamic routing. For details of how to configure routes, you must see the documentation that is delivered with your routing daemon (for example, zebra or gated).

**Adapter outage**

If outage of an adapter occurs, you must switch adapters.

**Procedure**

- Under static routing:
  1. Delete the route that was set previously.
  2. Create an alternative route to the virtual IP address.
- Under dynamic routing, see the documentation that is delivered with your routing daemon for details.

**Example of how to set up standard VIPA**

This example shows you how to configure VIPA under static routing, and how to switch adapters when an adapter outage occurs.

**About this task**

[Figure 56](#) shows the network adapter configuration that is used in the example.

---

**Diagram**

![Network Diagram](image)

**Procedure**

1. Define the real interfaces.

   ```
   [server]# ip addr add 10.1.0.2/16 dev enccw0.0.f500
   [server]# ip link set dev enccw0.0.f500 up
   [server]# ip addr add 10.2.0.2/16 dev enccw0.0.e400
   [server]# ip link set dev enccw0.0.e400 up
   ```

2. If the dummy component was not compiled into the kernel, ensure that the dummy module was loaded. If necessary, load it by issuing:

   ```
   [server]# modprobe dummy
   ```
3. Create a dummy interface with a virtual IP address 198.51.100.100 and a netmask 255.255.255.0:

```bash
[server]# ip addr add 198.51.100.100/24 dev dummy0
[server]# ip link set dev dummy0 up
```

4. Enable the network devices for this VIP so that it accepts packets for this IP address.
   - IPv4 example:
     ```bash
     [server]# qethconf vipa add 198.51.100.100 enccw0.0.f500
     qethconf: Added 198.51.100.100 to /sys/class/net/enccw0.0.f500/device/vipa/add4.
     qethconf: For verification please use "qethconf ipa list"
     [server]# qethconf vipa add 198.51.100.100 enccw0.0.e400
     qethconf: Added 198.51.100.100 to /sys/class/net/enccw0.0.e400/device/vipa/add4.
     qethconf: For verification please use "qethconf ipa list"
     ```
   - For IPv6, the address is specified in IPv6 format:
     ```bash
     [server]# qethconf vipa add 2002::1235:5678 enccw0.0.f500
     qethconf: Added 2002::1235:5678 to /sys/class/net/enccw0.0.f500/device/vipa/add6.
     qethconf: For verification please use "qethconf ipa list"
     [server]# qethconf vipa add 2002::1235:5678 enccw0.0.e400
     qethconf: Added 2002::1235:5678 to /sys/class/net/enccw0.0.e400/device/vipa/add6.
     qethconf: For verification please use "qethconf ipa list"
     ```

5. Ensure that the addresses are set:

6. Ensure that your service (such as the Apache web server) listens to the virtual IP address.

7. Set up a route to the virtual IP address (static routing) so that VIPA can be reached through the gateway with address 10.1.0.2.

```bash
[router]# ip route add 198.51.100.100 via 10.1.0.2
```

What to do next

Now assume that an adapter outage occurs. You must then:

1. Delete the previously created route.

```bash
[router]# ip route del 198.51.100.100
```

2. Create the alternative route to the virtual IP address.

```bash
[router]# ip route add 198.51.100.100 via 10.2.0.2
```

Source VIPA

Source VIPA is particularly suitable for high-performance environments. It selects one source address out of a range of source addresses when it replaces the source address of a socket.
Some operating system kernels cannot do load balancing among several connections with the same source and destination address over several interfaces. The solution is to use several source addresses.

To achieve load balancing, a policy must be selected in the policy section of the configuration file of Source VIPA (/etc/src_vipa.conf). In this policy section, you can also specify several source addresses that are used for one destination. Source VIPA then applies the source address selection according to the rules of the policy that is selected in the configuration file.

This Source VIPA solution does not affect kernel stability. Source VIPA is controlled by a configuration file that contains flexible rules for when to use Source VIPA based on destination IP address ranges.

You can use IPv6 or IPv4 addresses for Source VIPA. IPv6 addresses are supported as of Source VIPA 2.1.0.

**Setting up source VIPA**

To set up source VIPA, define your address ranges in the configuration file. Source VIPA is delivered as part of the s390utils package. Install the package as usual.

**Configuration**

With Source VIPA version 2.0.0 the configuration file changed: the policy section was added. The default configuration file is /etc/src_vipa.conf.

/etc/src_vipa.conf or the file pointed to by the environment variable SRC_VIPA_CONFIG_FILE, contains lines such as the following:

```plaintext
# comment
```

D1.D2.D3.D4/MASK specifies a range of destination addresses and the number of bits set in the subnet mask (MASK). As soon as a socket is opened and connected to these destination addresses and the application does not do an explicit bind to a source address, Source VIPA does a bind to one of the source addresses specified (S, T, [...]). It uses the policy that is selected in the configuration file to distribute the source addresses. See “Policies” on page 276 for available load distribution policies. Instead of IP addresses in dotted notation, host names can also be used and are resolved by using DNS.

You can use IPv6 or IPv4 IP addresses, but not both within a single rule in the configuration file. The following example shows an IPv6 configuration file with a random policy:

```plaintext
# IPv6
2221:11c3:0123:d9d8:05d5:5a44:724c:783b/64 random ed27:120:da42::
```

`.INADDR_ANY P1-P2 POLICY S1.S2.S3.S4` or `.INADDR_ANY P POLICY S1.S2.S3.S4` causes bind calls with .INADDR_ANY as a local address to be intercepted if the port the socket is bound to is between P1 and P2 (inclusive). In this case, .INADDR_ANY is replaced by one of the source addresses specified (S, T, [...]), which can be 0.0.0.0.

All .INADDR_ANY statements are read and evaluated in order of appearance. This method means that multiple .INADDR_ANY statements can be used to have bind calls...
intercepted for every port outside a certain range. This is useful, for example, for rlogin, which uses the bind command to bind to a local port but with \texttt{.INADDR\_ANY} as a source address to use automatic source address selection. See “Policies” for available load distribution policies.

The default behavior for all ports is that the kind of bind calls is not modified.

**Policies**

With Source VIPA Extensions, you provide a range of dummy source addresses for replacing the source addresses of a socket. The policy that is selected determines which method is used for selecting the source addresses from the range of dummy addresses.

**onevipa**

Only the first address of all source addresses specified is used as source address.

**random**

The source address that is used is selected randomly from all the specified source addresses.

**lrr (local round robin)**

The source address that is used is selected in a round robin manner from all the specified source addresses. The round robin takes place on a per-invocation base: each process is assigned the source addresses round robin independently from other processes.

**rr: ABC**

Stands for round robin and implements a global round robin over all Source VIPA instances that share a configuration file. All processes that use Source VIPA access an IPC shared memory segment to fulfil a global round robin algorithm. This shared memory segment is destroyed when the last running Source VIPA ends. However, if this process does not end gracefully (for example, is ended by a \texttt{kill} command), the shared memory segment (size: 4 bytes) can stay in the memory until it is removed by \texttt{ipcrm}. The tool \texttt{ipcs} can be used to display all IPC resources and to get the key or id used for \texttt{ipcrm}. ABC are UNIX permissions in octal writing (for example, 700) that are used to create the shared memory segment. Make this permission mask as restrictive as possible. A process that has access to this mask can cause an imbalance of the round robin distribution in the worst case.

**lc**

Attempts to balance the number of connections per source address. This policy always associates the socket with the VIPA that is least in use. If the policy cannot be parsed correctly, the policy is set to round robin per default.

**Enabling an application**

The command:

\texttt{src\_vipa.sh <application and parameters>}

enables the Source VIPA function for the application. The configuration file is read when the application is started. It is also possible to change the starter script and run multiple applications with different Source VIPA settings in separate files. To do this, define and export a \texttt{SRC\_VIPA\_CONFIG\_FILE} environment variable that points to the separate file before you start an application.
Note:
1. LD_PRELOAD security prevents setuid programs to be run under Source VIPA; programs of this kind can be run only when the real UID is 0. The ping utility is usually installed with setuid permissions.
2. The maximum number of VIPAs per destination is 8.

Example of how to set up source VIPA
An example shows how to set up source VIPA.

[Figure 57] shows a configuration where two applications with VIPA 198.51.100.100 and 198.51.100.200 are to be set up for Source VIPA with a local round robin policy.

The required entry in the Source VIPA configuration file is:

```
9.0.0.0/8 lrr 198.51.100.100 198.51.100.200
```

Scenario: Virtual LAN (VLAN) support
VLAN technology works according to IEEE Standard 802.1Q by logically segmenting the network into different broadcast domains. Thus packets are switched only between ports that are designated for the same VLAN.

By containing traffic that originates on a particular LAN to other LANs within the same VLAN, switched virtual networks avoid wasting bandwidth. Wasted bandwidth is a drawback inherent in traditional bridged/switched networks where packets are often forwarded to LANs that do not require them.

The qeth device driver for OSA-Express (QDIO) and HiperSockets supports priority tags as specified by IEEE Standard 802.1Q for both layer 2 and layer 3.

Introduction to VLANs
Use VLANs to increase traffic flow and reduce latency. With VLANs, you can organize your network by traffic patterns rather than by physical location.

In a conventional network topology, such as that shown in the following figure, devices communicate across LAN segments in different broadcast domains by
using routers. Although routers add latency by delaying transmission of data while they are using more of the data packet to determine destinations, they are preferable to building a single broadcast domain. A single domain can easily be flooded with traffic.

By organizing the network into VLANs by using Ethernet switches, distinct broadcast domains can be maintained without the latency that is introduced by multiple routers. As the following figure shows, a single router can provide the interfaces for all VLANs that appeared as separate LAN segments in the previous figure.

The following figure shows how VLANs can be organized logically, according to traffic flow, rather than being restricted by physical location. If workstations 1-3 communicate mainly with the small server, VLANs can be used to organize only these devices in a single broadcast domain that keeps broadcast traffic within the group. This setup reduces traffic both inside the domain and outside, on the rest of the network.
Configuring VLAN devices

Configure VLANs with the `ip link add` command. See the `ip-link` man page for details.

About this task

Information about the current VLAN configuration is available by listing the files in `/proc/net/vlan/*` with `cat` or `more`. For example:

```
# cat /proc/net/vlan/config
VLAN Dev name | VLAN ID
Name-Type: VLAN_NAME_TYPE_RAW_PLUS_VID_NO_PAD  bad_proto_recvd: 0
encw0.0.d300.100  100  encw0.0.d300
encw0.0.d300.200  200  encw0.0.d300
encw0.0.d300.300  300  encw0.0.d300

# cat /proc/net/vlan/enccw0.0.d300.300
enccw0.0.d300.300  VID: 300  REORDER_HDR: 1  dev->priv_flags: 1
   total frames received:  10914061
   total bytes received:  1291041929
Broadcast/Multicast Rcvd:  6
   total frames transmitted:  10471684
   total bytes transmitted:  4170258240
   total headroom inc:  0
   total encap on xmit:  10471684
Device: enccw0.0.d300
INGRESS priority mappings: 0:0 1:0 2:0 3:0 4:0 5:0 6:0 7:0
EGRESS priority Mappings:
```

Example: Creating two VLANs

VLANs are allocated in an existing interface that represents a physical Ethernet LAN.

The following example creates two VLANs, one with ID 3 and one with ID 5.

---

Figure 60. VLAN network organized for traffic flow
The `ip link add` commands added interfaces “enccw0.0.e400.3” and “enccw0.0.e400.5”, which you can then configure:

```
ip addr add 1.2.3.4/24 dev enccw0.0.e400.3
ip link set dev enccw0.0.e400.3 up
ip addr add 10.100.2.3/16 dev enccw0.0.e400.5
ip link set dev enccw0.0.e400.5 up
```

The traffic that flows out of enccw0.0.e400.3 is in the VLAN with ID=3. This traffic is not received by other stacks that listen to VLANs with ID=4.

The internal routing table ensures that every packet to 1.2.3.x goes out through enccw0.0.e400.3 and everything to 10.100.x.x through enccw0.0.e400.5. Traffic to 198.51.xxx.x flows through enccw0.0.e400 (without a VLAN tag).

To remove one of the VLAN interfaces:

```
ip link set dev enccw0.0.e400.3 down
ip link delete enccw0.0.e400.3 type vlan
```

---

**HiperSockets Network Concentrator**

You can configure a HiperSockets Network Concentrator on a QETH device in layer 3 mode.

**Before you begin:** The instructions that are given apply to IPv4 only. The HiperSockets Network Concentrator connector settings are available in layer 3 mode only.

The HiperSockets Network Concentrator connects systems to an external LAN within one IP subnet using HiperSockets. HiperSockets Network Concentrator connected systems look as if they were directly connected to the LAN. This simplification helps to reduce the complexity of network topologies that result from server consolidation.

Without changing the network setup, you can use HiperSockets Network Concentrator to port systems:

- From the LAN into an IBM Z Server environment
- From systems that are connected by a different HiperSockets Network Concentrator into an IBM Z Server environment

Thus, HiperSockets Network Concentrator helps to simplify network configuration and administration.

**Design**

A connector Linux system forwards traffic between the external OSA interface and one or more internal HiperSockets interfaces. The forwarding is done via IPv4 forwarding for unicast traffic and via a particular bridging code (xcec_bridge) for multicast traffic.
A script named `ip_watcher.pl` observes all IP addresses registered in the HiperSockets network and configures them as proxy ARP entries (see “Configuring a device for proxy ARP” on page 267) on the OSA interfaces. The script also establishes routes for all internal systems to enable IP forwarding between the interfaces.

All unicast packets that cannot be delivered in the HiperSockets network are handed over to the connector by HiperSockets. The connector also receives all multicast packets to bridge them.

**Setup**

The setup principles for configuring the HiperSockets Network Concentrator are as follows:

**leaf nodes**

The leaf nodes do not require a special setup. To attach them to the HiperSockets network, their setup should be as if they were directly attached to the LAN. They do not have to be Linux systems.

**connector systems**

In the following, HiperSockets Network Concentrator IP refers to the subnet of the LAN that is extended into the HiperSockets net.

- If you want to support forwarding of all packet types, define the OSA interface for traffic into the LAN as a multicast router (see “Setting up a Linux router in layer 3” on page 260).
- All HiperSockets interfaces that are involved must be set up as connectors: set the route4 attributes of the corresponding devices to “primary_connector” or to “secondary_connector”. Alternatively, you can add the OSA interface name to the start script as a parameter. This option results in HiperSockets Network Concentrator ignoring multicast packets, which are then not forwarded to the HiperSockets interfaces.
- IP forwarding must be enabled for the connector partition. Enable the forwarding either manually with the command

  ```bash
  sysctl -w net.ipv4.ip_forward=1
  ```

  Alternatively, you can enable IP forwarding in the `/etc/sysctl.conf` configuration file to activate IP forwarding for the connector partition automatically after booting.

  - The network routes for the HiperSockets interface must be removed. A network route for the HiperSockets Network Concentrator IP subnet must be established through the OSA interface. To establish a route, assign the IP address 0.0.0.0 to the HiperSockets interface. At the same time, assign an address used in the HiperSockets Network Concentrator IP subnet to the OSA interface. These assignments set up the network routes correctly for HiperSockets Network Concentrator.

  - To start HiperSockets Network Concentrator, run the script `start_hnc.sh`. You can specify an interface name as optional parameter. The interface name makes HiperSockets Network Concentrator use the specified interface to access the LAN. There is no multicast forwarding in that case.

  - To stop HiperSockets Network Concentrator, use the command `killall ip_watcher.pl` to remove changes that are caused by running HiperSockets Network Concentrator.
Availability setups

If a connector system fails during operation, it can simply be restarted. If all the startup commands are run automatically, it will instantaneously be operational again after booting. Two common availability setups are mentioned here:

**One connector partition and one monitoring system**

As soon as the monitoring system cannot reach the connector for a specific timeout (for example, 5 seconds), it restarts the connector. The connector itself monitors the monitoring system. If it detects (with a longer timeout than the monitoring system, for example, 15 seconds) a monitor system failure, it restarts the monitoring system.

**Two connector systems monitoring each other**

In this setup, there is an active and a passive system. As soon as the passive system detects a failure of the active connector, it takes over operation. To take over operation, it must reset the other system to release all OSA resources for the multicast_router operation. The failed system can then be restarted manually or automatically, depending on the configuration. The passive backup HiperSockets interface can either switch into primary_connector mode during the failover, or it can be set up as secondary_connector. A secondary_connector takes over the connecting function, as soon as there is no active primary_connector. This setup has a faster failover time than the first one.

**Hints**

- The MTU of the OSA and HiperSockets link should be of the same size. Otherwise, multicast packets that do not fit in the link’s MTU are discarded as there is no IP fragmentation for multicast bridging. Warnings are printed to /var/log/messages or a corresponding syslog destination.
- The script `ip_watcher.pl` prints error messages to the standard error descriptor of the process.
- `xcec-bridge` logs messages and errors to syslog. On Red Hat Enterprise Linux 7.5, you can find these messages in /var/log/messages.
- Registering all internal addresses with the OSA adapter can take several seconds for each address.
- To shut down the HiperSockets Network Concentrator function, issue `killall ip_watcher.pl`. This script removes all routing table and Proxy ARP entries added during the use of HiperSockets Network Concentrator.

**Note:**

1. Broadcast bridging is active only on OSA or HiperSockets hardware that can handle broadcast traffic without causing a bridge loop. If you see the message “Setting up broadcast echo filtering for ... failed” in the message log when you set the qeth device online, broadcast bridging is not available.
2. Unicast packets are routed by the common Linux IPv4 forwarding mechanisms. As bridging and forwarding are done at the IP Level, the IEEE 802.1q VLAN and the IPv6 protocol are not supported.

**Examples for setting up a network concentrator**

An example of a network environment with a network concentrator.

[Figure 61 on page 283](#) shows a network environment where a Linux instance C acts as a network concentrator that connects other operating system instances on a HiperSockets LAN to an external LAN.
Setup for the network concentrator C:

The HiperSockets interface enccw0.0.a1c0 (you can infer the interface name from the device bus-ID 0.0.a1c0) has IP address 10.20.30.51, and the netmask is 255.255.255.0. The default gateway is 10.20.30.1.

Issue:

```
# echo primary_connector > /sys/bus/ccwgroup/drivers/qeth/0.0.a1c0/route4
```

The OSA-Express CHPID in QDIO mode interface enca1c4 has IP address 10.20.30.11, and the netmask is 255.255.255.0. The default gateway is 10.20.30.1.

Issue:

```
# echo multicast_router > /sys/bus/ccwgroup/drivers/qeth/0.0.a1c4/route4
```

To enable IP forwarding issue:

```
# sysctl -w net.ipv4.ip_forward=1
```

To remove the network routes for the HiperSockets interface issue:

```
# ip route del 10.20.30/24
```

To start the HiperSockets network concentrator, run the script start_hsnc.sh. Issue:

```
# start_hsnc.sh &
```

Setup for G:

No special setup required. The HiperSockets interface has IP address 10.20.30.54, and the netmask is 255.255.255.0. The default gateway is 10.20.30.1.

Setup for workstation:

No special setup required. The network interface IP address is 10.20.30.120, and the netmask is 255.255.255.0. The default gateway is 10.20.30.1.
Figure 62 shows the example of Figure 61 on page 283 with an additional mainframe. On the second mainframe a Linux instance D acts as a HiperSockets network concentrator.

The configuration of C, G, and the workstation remain the same as for Figure 61 on page 283.

Setup for the network concentrator D:
The HiperSockets interface enccw0.0.a1c0 has IP address 0.0.0.0.
Assuming that the device bus-ID of the HiperSockets interface is 0.0.a1d0, issue:
```
# echo primary_connector > /sys/bus/ccwgroup/drivers/qeth/0.0.a1d0/route4
```

The OSA-Express CHPID in QDIO mode interface enca1d0 has IP address 10.20.30.50, and the netmask is 255.255.255.0. The default gateway is 10.20.30.1.

D is not configured as a multicast router, it therefore only forwards unicast packets.

To enable IP forwarding issue:
```
# sysct1 -w net.ipv4.ip_forward=1
```

Tip: See Red Hat Enterprise Linux 7.5 Installation Guide for information about using configuration files to automatically enable IP forwarding when Linux boots.
To start the HiperSockets network concentrator, run the script
start_hsnc.sh. Issue:

```
# start_hsnc.sh &
```

**Setup for H:**

No special setup required. The HiperSockets interface has IP address 10.20.30.55, and the netmask is 255.255.255.0. The default gateway is 10.20.30.1.

---

**Setting up for DHCP with IPv4**

For connections through an OSA-Express adapter in QDIO mode, the OSA-Express adapter offloads ARP, MAC header, and MAC address handling.

For information about MAC headers, see "MAC headers in layer 3 mode" on page 233.

Because a HiperSockets connection does not go out on a physical network, there are no ARP, MAC headers, and MAC addresses for packets in a HiperSockets LAN. The resulting problems for DHCP are the same in both cases and the fixes for connections through the OSA-Express adapter also apply to HiperSockets.

Dynamic Host Configuration Protocol (DHCP) is a TCP/IP protocol that allows clients to obtain IP network configuration information (including an IP address) from a central DHCP server. The DHCP server controls whether the address it provides to a client is allocated permanently or is leased temporarily. DHCP specifications are described by RFC 2131 “Dynamic Host Configuration Protocol” and RFC 2132 “DHCP options and BOOTP Vendor Extensions”, which are available at [www.ietf.org](http://www.ietf.org).

Two types of DHCP environments must be taken into account:

- DHCP that uses OSA-Express adapters in QDIO mode
- DHCP in a z/VM VSWITCH or guest LAN

For information about setting up DHCP for a Linux instance in a z/VM guest LAN environment, see Redpaper [Linux on IBM eServer™ zSeries and S/390: TCP/IP Broadcast on z/VM Guest LAN, REDP-3596 at www.ibm.com/redbooks](http://www.ibm.com/redbooks).

The programs `dhclient` and `dhcp` are examples of a DHCP client and a DHCP server you can use.

---

**Required options for using dhclient with layer3**

You must configure the DHCP client program `dhclient` to use it on Linux on Z with layer3.

- Run the DHCP client with an option that instructs the DHCP server to broadcast its response to the client.
  
  Because the OSA-Express adapter in QDIO mode forwards packets to Linux based on IP addresses, a DHCP client that requests an IP address cannot receive the response from the DHCP server without this option.
- Run the DHCP client with an option that specifies the client identifier string.
By default, the client uses the MAC address of the network interface. Hence, without this option, all Linux instances that share the OSA-Express adapter in QDIO mode would also have the same client identifier.

See the documentation for dhclient about how to select these options.

You need no special options for the DHCP server program, dhcp. You need no special options for using dhcp.

Setting up Linux as a LAN sniffer

You can set up a Linux instance to act as a LAN sniffer, for example, to make data on LAN traffic available to tools like tcpdump or Wireshark.

The LAN sniffer can be:

- A HiperSockets Network Traffic Analyzer for LAN traffic between LPARs
- A LAN sniffer for LAN traffic between z/VM guest virtual machines, for example, through a z/VM virtual switch (VSWITCH)

Setting up a HiperSockets network traffic analyzer

A HiperSockets network traffic analyzer (NTA) runs in an LPAR and monitors LAN traffic between LPARs.

Before you begin

- Your Linux instance must not be a z/VM guest.
- On the SE, the LPARs must be authorized for analyzing and being analyzed.

Tip: SE authorization changes for the HiperSockets network traffic analyzer require re-creating the device by ungrouping and regrouping (see "Removing a qeth group device" on page 242 and "Creating a qeth group device” on page 241). Do any authorization changes before you configure the NTA device.

- You need a traffic-dumping tool such as tcpdump.

About this task

HiperSockets NTA is available to trace both layer 3 and layer 2 network traffic, but the analyzing device itself must be configured as a layer 3 device. The analyzing device is a dedicated NTA device and cannot be used as a regular network interface.

Linux setup:

Ensure that the qeth device driver module was loaded.

Procedure

Perform the following steps:

1. Configure a HiperSockets interface dedicated to analyzing with the layer2 sysfs attribute set to 0 and the sniffer sysfs attribute set to 1. For example, assuming the HiperSockets interface is enccw0.0.a1c0 with device bus-ID 0.0.a1c0:

```
# znetconf -a a1c0 -o layer2=0 -o sniffer=1
```
The `znetconf` command also sets the device online. For more information about `znetconf`, see [“znetconf - List and configure network devices” on page 648](#). The qeth device driver automatically sets the buffer_count attribute to 128 for the analyzing device.

2. Activate the device (no IP address is needed):

   ```
   # ip link set enccw0.0.a1c0 up
   ```

3. Switch the interface into promiscuous mode:

   ```
   # tcpdump -i enccw0.0.a1c0
   ```

Results

The device is now set up as a HiperSockets network traffic analyzer.

**Hint:** A HiperSockets network traffic analyzer with no free empty inbound buffers might have to drop packets. Dropped packets are reflected in the “dropped counter” of the HiperSockets network traffic analyzer interface and reported by `tcpdump`.

**Example:**

```
# ip -s link show dev enccw0.0.a1c0
...
   RX: bytes packets errors dropped overrunt mcast
   223242  6789  0   5   0  176
...
# tcpdump -i enccw0.0.a1c0
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on enccw0.0.a1c0, link-type EN10MB (Ethernet), capture size 96 bytes
...
5 packets dropped by kernel
```

Setting up a z/VM guest LAN sniffer

You can set up a guest LAN sniffer on a virtual NIC that is coupled to a z/VM VSWITCH or guest LAN.

**Before you begin**

- You need class B authorization on z/VM.
- The Linux instance to be set up as a guest LAN sniffer must run as a guest of the same z/VM system as the guest LAN you want to investigate.

**About this task**

If a virtual switch connects to a VLAN that includes nodes outside the z/VM system, these external nodes are beyond the scope of the sniffer.

For information about VLANs and z/VM virtual switches, see z/VM Connectivity, SC24-6174.

**Procedure**

- Set up Linux.
  - Ensure that the qeth device driver is loaded.
- Set up z/VM.
Ensure that the z/VM guest virtual machine on which you want to set up the
guest LAN sniffer is authorized for the switch or guest LAN and for
promiscuous mode. For example, if your virtual NIC is coupled to a z/VM
virtual switch, perform the following steps on your z/VM system:

1. Check whether the z/VM guest virtual machine already has the requisite
authorizations. Enter a CP command of this form:

```
q vswitch <switchname> promisc
```

where `<switchname>` is the name of the virtual switch. If the output lists the
z/VM guest virtual machine as authorized for promiscuous mode, no further
setup is needed.

2. If the output from step [1] does not list the guest virtual machine, check if the
guest is authorized for the virtual switch. Enter a CP command of this form:

```
q vswitch <switchname> acc
```

where `<switchname>` is the name of the virtual switch.
If the output lists the z/VM guest virtual machine as authorized, you must
temporarily revoke the authorization for the switch before you can grant
authorization for promiscuous mode. Enter a CP command of this form:

```
set vswitch <switchname> revoke <userid>
```

where `<switchname>` is the name of the virtual switch and `<userid>` identifies
the z/VM guest virtual machine.

3. Authorize the Linux instance for the switch and for promiscuous mode.
Enter a CP command of this form:

```
set vswitch <switchname> grant <userid> promisc
```

where `<switchname>` is the name of the virtual switch and `<userid>` identifies
the z/VM guest virtual machine.

For details about the CP commands that are used here and for commands you
can use to check and assign authorizations for other types of guest LANs, see
z/VM CP Commands and Utilities Reference, SC24-6175.
Chapter 16. OSA-Express SNMP subagent support

The OSA-Express Simple Network Management Protocol (SNMP) subagent (osasnmpd) supports management information bases (MIBs) for the OSA-Express features.

The subagent supports OSA-Express features as shown in Table 34 on page 224.

This subagent capability through the OSA-Express features is also called Direct SNMP to distinguish it from another method of accessing OSA SNMP data through OSA/SF, a package for monitoring and managing OSA features that does not run on Linux.

To use the osasnmpd subagent, you need:

- An OSA-Express feature running in QDIO mode with the latest textual MIB file for the appropriate LIC level (recommended)
- The qeth device driver for OSA-Express (QDIO)
- The osasnmpd subagent from the s390utils RPM
- The net-snmp package delivered with Red Hat Enterprise Linux 6

What you should know about osasnmpd

The osasnmpd subagent requires a master agent to be installed on a Linux system.

You get the master agent from the net-snmp package. The subagent uses the Agent eXtensibility (AgentX) protocol to communicate with the master agent.

net-snmp is an open source project that is owned by the Open Source Development Network, Inc. (OSDN). For more information about net-snmp visit: net-snmp.sourceforge.net

When the master agent (snmpd) is started on a Linux system, it binds to a port (default 161) and awaits requests from SNMP management software. Subagents can connect to the master agent to support MIBs of special interest (for example, OSA-Express MIB). When the osasnmpd subagent is started, it retrieves the MIB objects of the OSA-Express features currently present on the Linux system. It then registers with the master agent the object IDs (OIDs) for which it can provide information.

An OID is a unique sequence of dot-separated numbers (for example, .1.3.6.1.4.1.2) that represents a particular information. OIDs form a hierarchical structure. The longer the OID, that is the more numbers it is made up of, the more specific is the information that is represented by the OID. For example, .1.3.6.1.4.1.2 represents all IBM-related network information while ..1.3.6.1.4.1.2.6.188 represents all OSA-Express-related information.

A MIB corresponds to a number of OIDs. MIBs provide information about their OIDs including textual representations the OIDs. For example, the textual representation of .1.3.6.1.4.1.2 is .iso.org.dod.internet.private.enterprises.ibm.
level, you must download the appropriate MIB file for the LIC level (see "Downloading the IBM OSA-Express MIB"). You do not need to update the subagent. Place the updated MIB file in a directory that is searched by the master agent.

![Diagram](image)

**Figure 63. OSA-Express SNMP agent flow**

**Figure 63** illustrates the interaction between the snmpd master agent and the osasnmpd subagent.

**Example:** This example shows the processes that run after the snmpd master agent and the osasnmpd subagent are started. When you start osasnmpd, a daemon called osasnmpd starts. In the example, PID 687 is the SNMP master agent and PID 729 is the OSA-Express SNMP subagent process:

```
ps -ef | grep snmp
```

<table>
<thead>
<tr>
<th>USER</th>
<th>PID</th>
</tr>
</thead>
<tbody>
<tr>
<td>root</td>
<td>687</td>
</tr>
<tr>
<td>root</td>
<td>729</td>
</tr>
</tbody>
</table>

When the master agent receives an SNMP request for an OID that is registered by a subagent, the master agent uses the subagent to collect any requested information and to perform any requested operations. The subagent returns any requested information to the master agent. Finally, the master agent returns the information to the originator of the request.

### Setting up osasnmpd

You must download the IBM OSA-Express MIB and configure access control before you use can the osasnmpd subagent.

### Downloading the IBM OSA-Express MIB

Keep your MIB file up to date by downloading the latest version.

### About this task

Perform the following steps to download the IBM OSA-Express MIB. The MIB file is valid only for hardware that supports the OSA-Express adapter.
Procedure

   A user ID and password are required. If you do not yet have one, you can apply for a user ID.
2. Sign in.
3. Select Library from the navigation area.
5. Follow the link for OSA-Express Direct SNMP MIB module.
6. Select and download the MIB for your LIC level.
7. Rename the MIB file to the name specified in the MIBs definition line and use the extension .txt.

   **Example:** If the definition line in the MIB looks like this:
   ```
   ==>IBM-OSA-MIB DEFINITIONS ::= BEGIN
   Rename the MIB to IBM-OSA-MIB.txt.
   ```

8. Place the MIB into /usr/share/snmp/mibs.
   If you want to use a different directory, be sure to specify the directory in the snmp.conf configuration file (see step 10 on page 294).

Results

You can now make the OID information from the MIB file available to the master agent. You can then use textual OIDs instead of numeric OIDs when you use master agent commands.

See also the FAQ (How do I add a MIB to the tools?) for the master agent package at [net-snmp.sourceforge.net/FAQ.html](http://net-snmp.sourceforge.net/FAQ.html)

Configuring access control

To start successfully, the subagent requires at least read access to the standard MIB-II on the local node.

About this task

During subagent startup or when network interfaces are added or removed, the subagent must query OIDs from the interfaces group of the standard MIB-II.

Given here is an example of how you can use the `snmpd.conf` and `snmpd.conf` configuration files to assign access rights with the View-Based Access Control Mechanism (VACM). The following access rights are assigned on the local node:

- General read access for the scope of the standard MIB-II
- Write access for the scope of the OSA-Express MIB
- Public local read access for the scope of the interfaces MIB

The example is intended for illustration purposes only. Depending on the security requirements of your installation, you might need to define your access differently. See the snmpd man page for more information about assigning access rights to snmpd.
Procedure

1. See the Red Hat Enterprise Linux 7.5 documentation to find out where you must place the snmpd.conf file. Some of the possible locations are:
   - /etc
   - /etc/snmp

2. Open snmpd.conf with your preferred text editor. There might be a sample in
   /usr/share/doc/packages/net-snmp/EXAMPLE.conf

3. Find the security name section and include a line of this form to map a community name to a security name:
   com2sec <security-name> <source> <community-name>

   where:
   
   <security-name>
   - is given access rights through further specifications within snmpd.conf.
   
   <source>
   - is the IP-address or DNS-name of the accessing system, typically a
     Network Management Station.
   
   <community-name>
   - is the community string used for basic SNMP password protection.

   Example:
   
   # sec.name source community
   com2sec osasec default osacom
   com2sec pubsec localhost public

4. Find the group section. Use the security name to define a group with different versions of the master agent for which you want to grant access rights.
   Include a line of this form for each master agent version:
   group <group-name> <security-model> <security-name>

   where:
   
   <group-name>
   - is a group name of your choice.
   
   <security-model>
   - is the security model of the SNMP version.
   
   <security-name>
   - is the same as in step 3

   Example:
   
   # groupName securityModel securityName
   group osagroup v1 osasec
   group osagroup v2c osasec
   group osagroup usm osasec
   group osasnmpd v2c pubsec

   Group “osasnmpd” with community “public” is required by osasnmpd to
determine the number of network interfaces.

5. Find the view section and define your views. A view is a subset of all OIDs.
   Include lines of this form:
   view <view-name> <included|excluded> <scope>

   where:
<view-name>
    is a view name of your choice.

<included|excluded>
    indicates whether the following scope is an inclusion or an exclusion statement.

<scope>
    specifies a subtree in the OID tree.

Example:

```plaintext
# name incl/excl subtree mask(optional)
view allview included .1
view osaview included .1.3.6.1.4.1.2
view ifmibview included interfaces
view ifmibview included system
```

View “allview” encompasses all OIDs while “osaview” is limited to IBM OIDs. The numeric OID provided for the subtree is equivalent to the textual OID “.iso.org.dod.internet.private.enterprises.ibm” View “ifmibview” is required by osasnmprd to determine the number of network interfaces.

Tip: Specifying the subtree with a numeric OID leads to better performance than using the corresponding textual OID.

6. Find the access section and define access rights. Include lines of this form:
   ```plaintext
   access <group-name> "" any noauth exact <read-view> <write-view> none
   ```

   where:
   - `<group-name>` is the group you defined in step 4 on page 292
   - `<read-view>` is a view for which you want to assign read-only rights.
   - `<write-view>` is a view for which you want to assign read-write rights.

Example:

```plaintext
# group context sec.model sec.level prefix read write notif
access osagroup "" any noauth exact allview osaview none
access osasnmprd "" v2c noauth exact ifmibview none none
```

The access line of the example gives read access to the “allview” view and write access to the “osaview”. The second access line gives read access to the “ifmibview”.

7. Also include the following line to enable the AgentX support:
   ```plaintext
   master agentx
   ```

   AgentX support is compiled into the net-snmp master agent.

8. Save and close snmpd.conf. Example of an snmpd.conf file:
   ```plaintext
   # name incl/excl subtree mask(optional)
   view allview included .1
   ```

**Tip:** See `man snmp.conf` for possible locations of `snmp.conf`.

10. Include a line of this form to specify the directory to be searched for MIBs:

```
mibdirs +<mib-path>
```

Example:

```
mibdirs +/usr/share/snmp/mibs
```

11. Include a line of this form to make the OSA-Express MIB available to the master agent:

```
mibs +<mib-name>
```

where `<mib-name>` is the stem of the MIB file name you assigned in "Downloading the IBM OSA-Express MIB” on page 290.

Example: `mibs +IBM-OSA-MIB`

12. Define defaults for the version and community to be used by the `snmp` commands. Add lines of this form:

```
defVersion <version>
defCommunity <community-name>
```

where `<version>` is the SNMP protocol version and `<community-name>` is the community you defined in step 3 on page 292.

Example:

```
defVersion 2c
defCommunity osacom
```

These default specifications simplify issuing master agent commands.

13. Save and close `~/.snmp/snmp.conf`.

---

**Working with the osasnmpd subagent**

Working with the osasnmpd subagent includes starting it, checking the log file, issuing queries, and stopping the subagent.

Working with osasnmpd comprises the following tasks:

- "Starting the osasnmpd subagent”
- "Checking the log file” on page 295
- “Issuing queries” on page 295
- “Stopping osasnmpd” on page 296

**Starting the osasnmpd subagent**

Use the `osasnmpd` command to start the osasnmpd subagent.
Procedure

After you download the s390utils-osasnmpd package and set up the osasnmpd subagent, start the subagent with the command:

```
# osasnmpd
```

The osasnmpd subagent, in turn, starts a daemon that is called osasnmpd. For command options see the `osasnmpd` command man page.

If you restart the master agent, you must also restart the subagent. When the master agent is started, it does not look for already running subagents. Any running subagents must also be restarted to be register with the master agent.

Checking the log file

Warnings and messages are written to the log file of either the master agent or the OSA-Express subagent. It is good practice to check these files at regular intervals.

Example

This example assumes that the default subagent log file is used. The lines in the log file show the messages after a successful OSA-Express subagent initialization.

```
# cat /var/log/osasnmpd.log
IBM OSA-E NET-SNMP 5.1.x subagent version 1.3.0
Jul 14 09:28:41 registered Toplevel OID .1.3.6.1.2.1.10.7.2.
Jul 14 09:28:41 registered Toplevel OID .1.3.6.1.4.1.2.6.188.1.1.
Jul 14 09:28:41 registered Toplevel OID .1.3.6.1.4.1.2.6.188.1.3.
Jul 14 09:28:41 registered Toplevel OID .1.3.6.1.4.1.2.6.188.1.4.
Jul 14 09:28:41 registered Toplevel OID .1.3.6.1.4.1.2.6.188.1.8.
OSA-E microcode level is 611 for interface enccw0.0.f500
Initialization of OSA-E subagent successful...
```

Issuing queries

You can issue queries against your SNMP setup.

About this task

Examples of what SNMP queries might look like are given here. For more comprehensive information about the master agent commands, see the `snmpcmd` man page.

The commands can use either numeric or textual OIDs. While the numeric OIDs might provide better performance, the textual OIDs are more meaningful and give a hint about which information is requested.

Examples

The query examples assume an interface, enccw0.0.f500, for which the CHPID is 6B. You can use the `lsqeth` command to find the mapping of interface names to CHPIDs.

- To list the ifIndex and interface description relation (on one line):

  ```
  # snmpget -v 2c -c osacom localhost interfaces.ifTable.ifEntry.ifDescr.6
  interfaces.ifTable.ifEntry.ifDescr.6 = enccw0.0.f500
  ```
Using this GET request you can see that enccw0.0.f500 has the ifIndex 6 assigned.

- To find the CHPID numbers for your OSA devices:

```sh
# snmpwalk -0 localhost .1.3.6.1.4.1.2.6.188.1.1.1.1
IBM-OSA-MIB::ibmOSAExpChannelNumber.6 = Hex-STRING: 00 6B
IBM-OSA-MIB::ibmOSAExpChannelNumber.7 = Hex-STRING: 00 7A
IBM-OSA-MIB::ibmOSAExpChannelNumber.8 = Hex-STRING: 00 7D
```

The first line of the command output, with index number 6, corresponds to CHPID 0x6B of the enccw0.0.f500 example. The example assumes that the community osacom is authorized as described in "Configuring access control" on page 291.

If you provided defaults for the SNMP version and the community (see step 12 on page 294), you can omit the `-v` and `-c` options:

```sh
# snmpwalk -0 localhost .1.3.6.1.4.1.2.6.188.1.1.1.1
IBM-OSA-MIB::ibmOSAExpChannelNumber.6 = Hex-STRING: 00 6B
IBM-OSA-MIB::ibmOSAExpChannelNumber.7 = Hex-STRING: 00 7A
IBM-OSA-MIB::ibmOSAExpChannelNumber.8 = Hex-STRING: 00 7D
```

You can obtain the same output by substituting the numeric OID .1.3.6.1.4.1.2.6.188.1.1.1.1 with its textual equivalent:

```sh
# snmpwalk -0 localhost ibmOsaExpChannelNumber
IBM-OSA-MIB::ibmOSAExpChannelNumber.6 = Hex-STRING: 00 6B
IBM-OSA-MIB::ibmOSAExpChannelNumber.7 = Hex-STRING: 00 7A
IBM-OSA-MIB::ibmOSAExpChannelNumber.8 = Hex-STRING: 00 7D
```

You can find the port type for the interface with index number 6:

```sh
# snmpwalk -0 localhost .1.3.6.1.4.1.2.6.188.1.4.1.2.6
IBM-OSA-MIB::ibmOsaExpEthPortType.6 = INTEGER: fastEthernet(81)
```

fastEthernet(81) corresponds to card type OSD_100.

Using the short form of the textual OID:

```sh
# snmpwalk -0 localhost ibmOsaExpEthPortType.6
IBM-OSA-MIB::ibmOsaExpEthPortType.6 = INTEGER: fastEthernet(81)
```

Specifying the index, 6 in the example, limits the output to the interface of interest.

**Stopping osasnmppd**

The subagent can be stopped by sending either a SIGINT or SIGTERM signal to the thread.

**About this task**

Avoid stopping the subagent with `kill -9` or with `kill -SIGKILL`. These commands do not allow the subagent to unregister the OSA-Express MIB objects from the SNMP master agent. This can cause problems when restarting the subagent.
If you saved the subagent PID to a file when you started it, you can consult this file for the PID. Otherwise, you can issue a `ps` command to find it out.

**Example**

The osasnpmd subagent starts a daemon that is called osasnpmd. To stop osasnpmd, issue the `kill` command for either the daemon or its PID:

```
# ps -ef | grep snmp
USER   PID   PS   PWR%  NICE  VIRT  RES  SHRN  S TTY      TIME  COMMAND
root   687    1   00:00 00:00:00 snmpd
root   729    1   00:00 00:00:00 osasnpmd
# killall osasnpmd
```
Chapter 17. LAN channel station device driver

The LAN channel station device driver (LCS device driver) supports Open Systems Adapters (OSA) features in non-QDIO mode.

Table 45 shows the supported OSA-Express features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>z14</th>
<th>z13 and z13s</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSA-Express6S</td>
<td>1000Base-T Ethernet</td>
<td>Not supported</td>
</tr>
<tr>
<td>OSA-Express5S</td>
<td>1000Base-T Ethernet</td>
<td>1000Base-T Ethernet</td>
</tr>
<tr>
<td>OSA-Express4S</td>
<td>1000Base-T Ethernet</td>
<td>1000Base-T Ethernet</td>
</tr>
<tr>
<td>OSA-Express3</td>
<td>Not supported</td>
<td>Not supported</td>
</tr>
</tbody>
</table>

Table 46. The LCS device driver supported OSA features on earlier mainframes

<table>
<thead>
<tr>
<th>Feature</th>
<th>zEC12 and zBC12</th>
<th>z196, z114, and System z10</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSA-Express4S</td>
<td>Not supported</td>
<td>Not supported</td>
</tr>
<tr>
<td>OSA-Express3</td>
<td>Not supported</td>
<td>1000Base-T Ethernet</td>
</tr>
<tr>
<td>OSA-Express2</td>
<td>1000Base-T Ethernet</td>
<td>1000Base-T Ethernet</td>
</tr>
<tr>
<td>OSA-Express</td>
<td>1000Base-T Ethernet</td>
<td>Not supported</td>
</tr>
</tbody>
</table>

The LCS device driver supports automatic detection of Ethernet connections. The LCS device driver can be used for Internet Protocol, version 4 (IPv4) only.

What you should know about LCS

Interface names are assigned to LCS group devices, which map to subchannels and their corresponding device numbers and device bus-IDs.

LCS group devices

The LCS device driver requires two I/O subchannels for each LCS interface, a read subchannel and a write subchannel. The corresponding bus IDs must be configured for control unit type 3088.

The device bus-IDs that correspond to the subchannel pair are grouped as one LCS group device. The following rules apply for the device bus-IDs:
read must be even.
write must be the device bus-ID of the read subchannel plus one.

Setting up the LCS device driver

There are no module parameters for the LCS device driver.

You must load the lcs module before you can work with LCS devices. Load the lcs module with the `modprobe` command to ensure that any other required modules are loaded in the correct order:

```
# modprobe lcs
```

Working with LCS devices

Working with LCS devices includes tasks such as creating an LCS group device, specifying a timeout, or activating an interface.

- “Creating an LCS group device”
- “Removing an LCS group device” on page 301
- “Specifying a timeout for LCS LAN commands” on page 302
- “Setting an LCS group device online or offline” on page 302
- “Activating and deactivating an interface” on page 303
- “Recovering an LCS group device” on page 304

Most of these tasks involve writing to and reading from device attributes in sysfs. Using attributes is useful on a running system where you want to make dynamic changes. If you want to make persistent changes across IPLs, use the interface configuration files. Network configuration parameters are defined in `/etc/sysconfig/network-scripts/ifcfg-<if_name>`. An example of how to define an LCS device persistently is in Red Hat Enterprise Linux 7.5 Installation Guide. For a general discussion of network configuration files, see Red Hat Enterprise Linux 7.5 System Administrator’s Guide.

Creating an LCS group device

Use the group attribute to create an LCS group device.

Before you begin

You must know the device bus-IDs that correspond to the read and write subchannel of your OSA card. The subchannel is defined in the IOCDS of your mainframe.

Procedure

To define an LCS group device, write the device bus-IDs of the subchannel pair to `/sys/bus/ccwgroup/drivers/lcs/group`. Issue a command of this form:

```
# echo <read_device_bus_id>,<write_device_bus_id> > /sys/bus/ccwgroup/drivers/lcs/group
```
Results

The lcs device driver uses the device bus-ID of the read subchannel to create a directory for a group device:

```
/sys/bus/ccwgroup/drivers/lcs/<read_device_bus_id>
```

This directory contains a number of attributes that determine the settings of the LCS group device. The following sections describe how to use these attributes to configure an LCS group device.

Example

Assuming that 0.0.d000 is the device bus-ID that corresponds to a read subchannel:

```
# echo 0.0.d000,0.0.d001 > /sys/bus/ccwgroup/drivers/lcs/group
```

This command results in the creation of the following directories in sysfs:

- /sys/bus/ccwgroup/drivers/lcs/0.0.d000
- /sys/bus/ccwgroup/devices/0.0.d000
- /sys/devices/lcs/0.0.d000

Note: When the device subchannels are added, device types 3088/08 and 3088/1f can be assigned to either the CTCM or the LCS device driver.

To check which devices are assigned to which device driver, issue the following commands:

```
# ls -l /sys/bus/ccw/drivers/ctcm
# ls -l /sys/bus/ccw/drivers/lcs
```

To change a faulty assignment, use the unbind and bind attributes of the device. For example, to change the assignment for device bus-IDs 0.0.2000 and 0.0.2001 issue the following commands:

```
# echo 0.0.2000 > /sys/bus/ccw/drivers/ctcm/unbind
# echo 0.0.2000 > /sys/bus/ccw/drivers/lcs/bind
# echo 0.0.2001 > /sys/bus/ccw/drivers/ctcm/unbind
# echo 0.0.2001 > /sys/bus/ccw/drivers/lcs/bind
```

Removing an LCS group device

Use the ungroup attribute to remove an LCS group device.

Before you begin

The device must be set offline before you can remove it.

Procedure

To remove an LCS group device, write 1 to the ungroup attribute. Issue a command of the form:

```
echo 1 > /sys/bus/ccwgroup/drivers/lcs/<device_bus_id>/ungroup
```
Example

This command removes device 0.0.d000:

```
echo 1 > /sys/bus/ccwgroup/drivers/lcs/0.0.d000/ungroup
```

Specifying a timeout for LCS LAN commands

Use the `lancmd_timeout` attribute to set a timeout for an LCS LAN command.

About this task

You can specify a timeout for the interval that the LCS device driver waits for a reply after issuing a LAN command to the LAN adapter. For older hardware, the replies can take a longer time. The default is 5 s.

Procedure

To set a timeout, issue a command of this form:

```
# echo <timeout> > /sys/bus/ccwgroup/drivers/lcs/<device_bus_id>/lancmd_timeout
```

where `<timeout>` is the timeout interval in seconds in the range 1 - 60.

Example

In this example, the timeout for a device 0.0.d000 is set to 10 s.

```
# echo 10 > /sys/bus/ccwgroup/drivers/lcs/0.0.d000/lancmd_timeout
```

Setting an LCS group device online or offline

Use the `online` device group attribute to set an LCS device online or offline.

About this task

Setting a device online associates it with an interface name. Setting the device offline preserves the interface name.

Read `/var/log/messages` or issue `dmesg` to determine the assigned interface name. You must know the interface name to activate the network interface.

For each online interface, there is a symbolic link of the form `/sys/class/net/<interface_name>/device` in sysfs. You can confirm that you found the correct interface name by reading the link.

Procedure

To set an LCS group device online, set the online device group attribute to 1. To set an LCS group device offline, set the online device group attribute to 0. Issue a command of this form:

```
# echo <flag> > /sys/bus/ccwgroup/drivers/lcs/<device_bus_id>/online
```
Example

To set an LCS device with bus ID 0.0.d000 online issue:

```
# echo 1 > /sys/bus/ccwgroup/drivers/lcs/0.0.d000/online
# dmesg
...
lcs: LCS device encd000 without IPv6 support
lcs: LCS device encd000 with Multicast support
...
```

The interface name that was assigned to the LCS group device in the example is encd000. To confirm that this name is correct for the group device issue:

```
# readlink /sys/class/net/encd000/device
../../../devices/lcs/0.0.d000
```

If an error occurs when you set the device online, ensure that the physical connection from the port to the network is in place. If the error persists, note the return code from the error message and contact IBM support.

To set the device offline issue:

```
# echo 0 > /sys/bus/ccwgroup/drivers/lcs/0.0.d000/online
```

Activating and deactivating an interface

Use the `ip` command or equivalent to activate or deactivate an interface.

About this task

Before you can activate an interface, you must set the group device online and find out the interface name that is assigned by the LCS device driver. See “Setting an LCS group device online or offline” on page 302.

You activate or deactivate network devices with `ip` or an equivalent command. For details of the `ip` command, see the `ip` man page.

Examples

- This example activates an Ethernet interface:
  ```
  # ip addr add 192.168.100.10/24 dev enccw0.0.f500
  # ip link set dev enccw0.0.f500 up
  ```

- This example deactivates the Ethernet interface:
  ```
  # ip link set dev enccw0.0.f500 down
  ```

- This example reactivates an interface that was already activated and subsequently deactivated:
  ```
  # ip link set dev enccw0.0.f500 up
  ```
Recovering an LCS group device

You can use the recover attribute of an LCS group device to recover it in case of failure. For example, error messages in /var/log/messages might inform you of a malfunctioning device.

Procedure

Issue a command of the form:

```
# echo 1 > /sys/bus/ccwgroup/drivers/lcs/<device_bus_id>/recover
```

Example

```
# echo 1 > /sys/bus/ccwgroup/drivers/lcs/0.0.d100/recover
```
Chapter 18. CTCM device driver

The CTCM device driver provides Channel-to-Channel (CTC) connections and CTC-based Multi-Path Channel (MPC) connections. The CTCM device driver is required by Communications Server for Linux.

**Deprecated connection type:** CTC connections are deprecated. Do not use for new network setups.

CTC connections are high-speed point-to-point connections between two mainframe operating system instances.

Communications Server for Linux uses MPC connections to connect Red Hat Enterprise Linux 7.5 to VTAM® on traditional mainframe operating systems.

**Features**

The CTCM device driver provides different kinds of CTC connections between mainframes, z/VM guests, and LPARs.

The CTCM device driver provides:

- MPC connections to VTAM on traditional mainframe operating systems.
- ESCON or FICON CTC connections (standard CTC and basic CTC) between mainframes in basic mode, LPARs or z/VM guests.
  
  For more information about FICON, see Redpaper *FICON CTC Implementation*, REDP-0158.
- Virtual CTCA connections between guests of the same z/VM system.
- CTC connections to other Linux instances or other mainframe operating systems.

**What you should know about CTCM**

The CTCM device driver assigns network interface names to CTCM group devices.

**CTCM group devices**

The CTCM device driver requires two I/O subchannels for each interface, a read subchannel and a write subchannel.

The CTCM device driver requires two I/O subchannels for each interface, a read subchannel and a write subchannel (see Figure 65 on page 306). The device bus-IDs that correspond to the two subchannels must be configured for control unit type 3088.
The device bus-IDs that correspond to the subchannel pair are grouped as one CTCM group device. There are no constraints on the device bus-IDs of read subchannel and write subchannel. In particular, it is possible to group non-consecutive device bus-IDs.

On the communication-peer operating system instance, read and write subchannels are reversed. That is, the write subchannel of the local interface is connected to the read subchannel of the remote interface and vice versa.

Depending on the protocol, the interfaces can be CTC interfaces or MPC interfaces. MPC interfaces are used by Communications Server for Linux and connect to peer interfaces that run under VTAM. For more information about Communications Server for Linux and on using MPC connections, go to [www.ibm.com/software/network/commserver/linux](http://www.ibm.com/software/network/commserver/linux).

### Interface names assigned by the CTCM device driver

When a CTCM group device is set online, the CTCM device driver automatically assigns an interface name to it. The interface name depends on the protocol.

If the protocol is set to 4, you get an MPC connection and the interface names are of the form `mpc<n>`.

If the protocol is set to 0, 1, or 3, you get a CTC connection and the interface name is of the form `ctc<n>`.

<n> is an integer that identifies the device. When the first device is set online it is assigned 0, the second is assigned 1, the third 2, and so on. The devices are counted separately for CTC and MPC.

### Network connections

If your CTC connection is to a router or z/VM TCP/IP service machine, you can connect CTC interfaces to an external network.

[Figure 66 on page 307](#) shows a CTC interface that is connected to a network.
Setting up the CTCM device driver

There are no module parameters for the CTCM device driver. Load the ctcm module before you use it.

Load the ctcm module with the `modprobe` command to ensure that any other required modules are loaded:

```
# modprobe ctcm
```

Working with CTCM devices

When you work with CTCM devices you might create a CTCM group device, set the protocol, and activate an interface.

The following sections describe typical tasks that you need when you work with CTCM devices.

- "Creating a CTCM group device"
- "Removing a CTCM group device" on page 308
- "Displaying the channel type" on page 309
- "Setting the protocol" on page 309
- "Setting a device online or offline" on page 310
- "Setting the maximum buffer size" on page 311 (CTC only)
- "Activating and deactivating a CTC interface" on page 311 (CTC only)
- "Recovering a lost CTC connection" on page 313 (CTC only)

See the Communications Server for Linux documentation for information about configuring and activating MPC interfaces.

Creating a CTCM group device

Use the group attribute to create a CTCM group device.

Before you begin

You must know the device bus-IDs that correspond to the local read and write subchannel of your CTCM connection as defined in your IOCDS.
**Procedure**

To define a CTCM group device, write the device bus-IDs of the subchannel pair to `/sys/bus/ccwgroup/drivers/ctcm/group`. Issue a command of this form:

```
# echo <read_device_bus_id>,<write_device_bus_id> > /sys/bus/ccwgroup/drivers/ctcm/group
```

**Results**

The CTCM device driver uses the device bus-ID of the read subchannel to create a directory for a group device:

`/sys/bus/ccwgroup/drivers/ctcm/<read_device_bus_id>`

This directory contains a number of attributes that determine the settings of the CTCM group device.

**Example**

Assuming that device bus-ID 0.0.2000 corresponds to a read subchannel:

```
# echo 0.0.2000,0.0.2001 > /sys/bus/ccwgroup/drivers/ctcm/group
```

This command results in the creation of the following directories in sysfs:

- `/sys/bus/ccwgroup/drivers/ctcm/0.0.2000`
- `/sys/bus/ccwgroup/devices/0.0.2000`
- `/sys/devices/ctcm/0.0.2000`

**Note:** When the device subchannels are added, device types 3088/08 and 3088/1f can be assigned to either the CTCM or the LCS device driver.

To check which devices are assigned to which device driver, issue the following commands:

```
# ls -l /sys/bus/ccw/drivers/ctcm
# ls -l /sys/bus/ccw/drivers/lcs
```

To change a faulty assignment, use the unbind and bind attributes of the device. For example, to change the assignment for device bus-IDs 0.0.2000 and 0.0.2001 issue the following commands:

```
# echo 0.0.2000 > /sys/bus/ccw/drivers/lcs/unbind
# echo 0.0.2000 > /sys/bus/ccw/drivers/ctcm/bind
# echo 0.0.2001 > /sys/bus/ccw/drivers/lcs/unbind
# echo 0.0.2001 > /sys/bus/ccw/drivers/ctcm/bind
```

**Removing a CTCM group device**

Use the ungroup attribute to remove a CTCM group device.

**Before you begin**

The device must be set offline before you can remove it.
Procedure

To remove a CTCM group device, write 1 to the ungroup attribute. Issue a command of the form:

```
# echo 1 > /sys/bus/ccwgroup/drivers/ctcm/<device_bus_id>/ungroup
```

Example

This command removes device 0.0.2000:

```
echo 1 > /sys/bus/ccwgroup/drivers/ctcm/0.0.2000/ungroup
```

Displaying the channel type

Use the type attribute to display the channel type of a CTCM group device.

Procedure

Issue a command of this form to display the channel type of a CTCM group device:

```
# cat /sys/bus/ccwgroup/drivers/ctcm/<device_bus_id>/type
```

where `<device_bus_id>` is the device bus-ID that corresponds to the CTCM read channel. Possible values are: CTC/A, ESCON, and FICON.

Example

In this example, the channel type is displayed for a CTCM group device with device bus-ID 0.0.f000:

```
# cat /sys/bus/ccwgroup/drivers/ctcm/0.0.f000/type
ESCON
```

Setting the protocol

Use the protocol attribute to set the protocol.

Before you begin

The device must be offline while you set the protocol.

About this task

The type of interface depends on the protocol. Protocol 4 results in MPC interfaces with interface names `mpc<n>`. Protocols 0, 1, or 3 result in CTC interfaces with interface names of the form `ctc<n>`.

To choose a protocol, set the protocol attribute to one of the following values:

0  This protocol provides compatibility with peers other than z/OS, for example, a z/VM TCP service machine. This value is the default.
1  This protocol provides enhanced package checking for Linux peers.
This protocol provides for compatibility with z/OS peers.

This protocol provides for MPC connections to VTAM on traditional mainframe operating systems.

**Procedure**

Issue a command of this form:

```
# echo <value> > /sys/bus/ccwgroup/drivers/ctcm/<device_bus_id>/protocol
```

**Example**

In this example, the protocol is set for a CTCM group device 0.0.2000:

```
# echo 4 > /sys/bus/ccwgroup/drivers/ctcm/0.0.2000/protocol
```

**Setting a device online or offline**

Use the `online` device group attribute to set a CTCM device online or offline.

**About this task**

Setting a group device online associates it with an interface name. Setting the group device offline and back online with the same protocol preserves the association with the interface name. If you change the protocol before you set the group device back online, the interface name can change as described in "Interface names assigned by the CTCM device driver" on page 306.

You must know the interface name to access the CTCM group device. To determine the assigned interface name, use the `znetconf -c` command. For each online interface, the interface name is shown in the Name column. Alternatively, to determine the assigned interface name issue a command of the form:

```
# ls /sys/devices/ctcm/<device_bus_id>/net/
```

For each online interface, there is a symbolic link of the form `/sys/class/net/<interface_name>/device` in sysfs. You can confirm that you found the correct interface name by reading the link.

**Procedure**

To set a CTCM group device online, set the online device group attribute to 1. To set a CTCM group device offline, set the online device group attribute to 0. Issue a command of this form:

```
# echo <flag> > /sys/bus/ccwgroup/drivers/ctcm/<device_bus_id>/online
```

**Example**

To set a CTCM device with bus ID 0.0.2000 online issue:

```
# echo 1 > /sys/bus/ccwgroup/drivers/ctcm/0.0.2000/online
```
To determine the interface name issue:

```
# znetconf -c
```

<table>
<thead>
<tr>
<th>Device IDs</th>
<th>Type</th>
<th>Card Type</th>
<th>ChPID</th>
<th>Drv.</th>
<th>Name</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.2000,0.0.2001</td>
<td>3088/08</td>
<td>CTC/A</td>
<td>ctcmslccw0.0.2000</td>
<td>online</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

or

```
# ls /sys/devices/ctcm/0.0.2000/net/
slccw0.0.2000
```

To set group device 0.0.2000 offline issue:

```
# echo 0 > /sys/bus/ccwgroup/drivers/ctcm/0.0.2000/online
```

### Setting the maximum buffer size

Use the buffer device group attribute to set a maximum buffer size for a CTCM group device.

**Before you begin**

- Set the maximum buffer size for CTC interfaces only. MPC interfaces automatically use the highest possible maximum buffer size.
- The device must be online when you set the buffer size.

**About this task**

You can set the maximum buffer size for a CTC interface. The permissible range of values depends on the MTU settings. It must be in the range \( <\text{minimum MTU + header size}> \) to \( <\text{maximum MTU + header size}> \). The header space is typically 8 byte. The default for the maximum buffer size is 32768 byte (32 KB).

Changing the buffer size is accompanied by an MTU size change to the value \( <\text{buffer size - header size}> \).

**Procedure**

To set the maximum buffer size, issue a command of this form:

```
# echo <value> > /sys/bus/ccwgroup/drivers/ctcm/<device_bus_id>/buffer
```

where \(<value>\) is the number of bytes you want to set. If you specify a value outside the valid range, the command is ignored.

**Example**

In this example, the maximum buffer size of a CTCM group device 0.0.f000 is set to 16384 byte.

```
# echo 16384 > /sys/bus/ccwgroup/drivers/ctcm/0.0.f000(buffer
```

### Activating and deactivating a CTC interface

Use `ip` or an equivalent command to activate or deactivate an interface.
Before you begin

- Activate and deactivate a CTC interfaces only. For information about how to activate MPC interfaces, see the Communications Server for Linux documentation.
- You must know the interface name. See "Setting a device online or offline" on page 310.

About this task

Syntax for setting an IP address for a CTC interface with the ip command

```
ip address add <ip_address> dev <interface>
peer <peer_ip_address>
```

Syntax for activating a CTC interface with the ip command

```
ip link set dev <interface> up
mtu 32760
mtu <max_transfer_unit>
```

Where:

- `<interface>`
  - is the interface name that was assigned when the CTCM group device was set online.
- `<ip_address>`
  - is the IP address that you want to assign to the interface.
- `<peer_ip_address>`
  - is the IP address of the remote side.
- `<max_transfer_unit>`
  - is the size of the largest IP packet that might be transmitted. Be sure to use the same MTU size on both sides of the connection. The MTU must be in the range of 576 byte to 65,536 byte (64 KB).

Syntax for deactivating a CTC interface with the ip command

```
ip link set dev <interface> down
```

Where:
<interface>

is the interface name that was assigned when the CTCM group device was set online.

**Procedure**

- Use `ip` or an equivalent command to activate the interface.
- To deactivate an interface, issue a command of this form:

  ```
  # ip link set dev <interface> down
  ```

**Examples**

- This example activates a CTC interface slccw0.0.2000 with an IP address 10.0.51.3 for a peer with address 10.0.50.1 and an MTU of 32760.

  ```
  # ip addr add 10.0.51.3 dev slccw0.0.2000 peer 10.0.50.1
  # ip link set dev slccw0.0.2000 up mtu 32760
  ```

- This example deactivates slccw0.0.2000:

  ```
  # ip link set dev slccw0.0.2000 down
  ```

**Recovering a lost CTC connection**

If one side of a CTC connection crashes, you cannot simply reconnect after a reboot. You must also deactivate the interface of the peer of the crashed side.

**Before you begin**

These instructions apply to CTC interfaces only.

**Procedure**

Proceed as follows to recover a lost CTC connection:

1. Reboot the crashed side.
2. Deactivate the interface on the peer. See “Activating and deactivating a CTC interface” on page 311.
3. Activate the interface on the crashed side and on the peer. For details, see “Activating and deactivating a CTC interface” on page 311.

   If the connection is between a Linux instance and a non-Linux instance, activate the interface on the Linux instance first. Otherwise, you can activate the interfaces in any order.

**Results**

If the CTC connection is uncoupled, you must couple it again and reconfigure the interface of both peers with the `ip` command. See “Activating and deactivating a CTC interface” on page 311.
CTCM scenarios

Typical use cases of CTC connections include connecting to a peer in a different LPAR and connecting Linux instances running as z/VM guests to each other.

This section provides some typical scenarios for CTC connections:

- “Connecting to a peer in a different LPAR”
- “Connecting Linux on z/VM to another guest of the same z/VM system” on page 315

Connecting to a peer in a different LPAR

A Linux instance and a peer run in LPAR mode on the same or on different mainframes and are to be connected with a CTC FICON or CTC ESCON network interface.

Assumptions:

- Locally, the read and write channels have been configured for type 3088 and use device bus-IDs 0.0.f008 and 0.0.f009.
- IP address 10.0.50.4 is to be used locally and 10.0.50.5 for the peer.

Figure 67 illustrates a CTC setup with a peer in a different LPAR.

![Figure 67. CTC scenario with peer in a different LPAR](image)

**Procedure**

1. Create a CTCM group device. Issue:
   
   ```bash
   # echo 0.0.f008,0.0.f009 > /sys/bus/ccwgroup/drivers/ctcm/group
   ```

2. Confirm that the device uses CTC FICON or CTC ESCON:
   
   ```bash
   # cat /sys/bus/ccwgroup/drivers/ctcm/0.0.f008/type
   ESCON
   ```

   In this example, ESCON is used. You would proceed the same for FICON.

3. Select a protocol. The choice depends on the peer.
<table>
<thead>
<tr>
<th>If the peer is ...</th>
<th>Choose ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linux</td>
<td>1</td>
</tr>
<tr>
<td>z/OS or OS/390</td>
<td>3</td>
</tr>
<tr>
<td>Any other operating system</td>
<td>0</td>
</tr>
</tbody>
</table>

Assuming that the peer is Linux:

```
# echo 1 > /sys/bus/ccwgroup/drivers/ctcm/0.0.f008/protocol
```

4. Set the CTCM group device online and find out the assigned interface name:

```
# echo 1 > /sys/bus/ccwgroup/drivers/ctcm/0.0.f008/online
# ls /sys/devices/ctcm/0.0.f008/net/
enccw0.0.f008
```

In the example, the interface name is enccw0.0.f008.

5. Assure that the peer interface is configured.

6. Activate the interface locally and on the peer. If you are connecting two Linux instances, either instance can be activated first. If the peer is not Linux, activate the interface on Linux first. To activate the local interface:

```
# ip addr add 10.0.50.4 dev enccw0.0.f008 peer 10.0.50.5
# ip link set dev enccw0.0.f008 up
```

**Connecting Linux on z/VM to another guest of the same z/VM system**

A virtual CTCA connection is to be set up between an instance of Linux on z/VM and another guest of the same z/VM system.

**Assumptions:**

- The guest ID of the peer is “guestp”.
- A separate subnet has been obtained from the TCP/IP network administrator. The Linux instance will use IP address 10.0.100.100 and the peer will use IP address 10.0.100.101.

Figure 68 illustrates a CTC setup with a peer in the same z/VM.

![Figure 68. CTC scenario with peer in the same z/VM](image-url)
Procedure

1. Define two virtual channels to your user ID. The channels can be defined in the z/VM user directory using directory control SPECIAL statements, for example:
   special f004 cbecca
   special f005 cbecca

   Alternatively, you can use the CP commands:
   
   define cbecca as f004
   define cbecca as f005

2. Assure that the peer interface is configured.
3. Connect the virtual channels. Assuming that the read channel on the peer corresponds to device number 0xf010 and the write channel to 0xf011 issue:
   
   couple f004 to guestp f011
   couple f005 to guestp f010

   Be sure that you couple the read channel to the peers write channel and vice versa.
4. From your booted Linux instance, create a CTCM group device. Issue:
   
   # echo 0.0.f004,0.0.f005 > /sys/bus/ccwgroup/drivers/ctcm/group

5. Confirm that the group device is a virtual CTCA device:
   
   # cat /sys/bus/ccwgroup/drivers/ctcm/0.0.f004/type
   CTC/A

6. Select a protocol. The choice depends on the peer.

<table>
<thead>
<tr>
<th>If the peer is ...</th>
<th>Choose ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linux</td>
<td>1</td>
</tr>
<tr>
<td>z/OS or OS/390®</td>
<td>3</td>
</tr>
<tr>
<td>Any other operating system</td>
<td>0</td>
</tr>
</tbody>
</table>

   Assuming that the peer is Linux:
   
   # echo 1 > /sys/bus/ccwgroup/drivers/ctcm/0.0.f004/protocol

7. Set the CTCM group device online and find out the assigned interface name:
   
   # echo 1 > /sys/bus/ccwgroup/drivers/ctcm/0.0.f004/online
   # ls /sys/devices/ctcm/0.0.f004/net/
   enccw0.0.f004

   In the example, the interface name is enccw0.0.f004.
8. Activate the interface locally and on the peer. If you are connecting two Linux instances, either can be activated first. If the peer is not Linux, activate the local interface first. To activate the local interface:
   
   # ip addr add 10.0.100.100 dev enccw0.0.f004 peer 10.0.100.101
   # ip link set dev enccw0.0.f004 up
Be sure that the MTU on both sides of the connection is the same. If necessary change the default MTU (see “Activating and deactivating a CTC interface” on page 311).

9. Ensure that the buffer size on both sides of the connection is the same. For the Linux side see “Setting the maximum buffer size” on page 311. If the peer is not Linux, see the operating system documentation of the peer.
Chapter 19. AF_IUCV address family support

The AF_IUCV address family provides an addressing mode for communications between applications that run on Z.

This addressing mode can be used for connections through real HiperSockets and through the z/VM Inter-User Communication Vehicle (IUCV).

Support for AF_IUCV based connections through real HiperSockets requires Completion Queue Support.

HiperSockets devices facilitate connections between applications across LPARs within an IBM Z. In particular, an application that runs on an instance of Linux on Z can communicate with:

- Itself
- Other applications that run on the same Linux instance
- An application on an instance of Linux on Z in another LPAR

IUCV facilitates connections between applications across z/VM guest virtual machines within a z/VM system. In particular, an application that runs on Linux on z/VM can communicate with:

- Itself
- Other applications that run on the same Linux instance
- Applications running on other instances of Linux on z/VM, within the same z/VM system
- Applications running on a z/VM guest other than Linux, within the same z/VM system
- The z/VM control program (CP)

The AF_IUCV address family supports stream-oriented sockets (SOCK_STREAM) and connection-oriented datagram sockets (SOCK_SEQPACKET). Stream-oriented sockets can fragment data over several packets. Sockets of type SOCK_SEQPACKET always map a particular socket write or read operation to a single packet.

Features

The AF_IUCV address family provides socket connections for HiperSockets and IUCV.

For all instances of Linux on Z, the AF_IUCV address family provides the following features:

- Multiple outgoing socket connections for real HiperSockets
- Multiple incoming socket connections for real HiperSockets

For instances of Linux on z/VM, the AF_IUCV address family also provides the following features:

- Multiple outgoing socket connections for IUCV
- Multiple incoming socket connections for IUCV
• Socket communication with applications that use the CMS AF_IUCV support

### Setting up the AF_IUCV address family support

You must authorize your z/VM guest virtual machine and load those components that were compiled as separate modules.

There are no module parameters for the AF_IUCV address family support.

### Setting up HiperSockets devices for AF_IUCV addressing

In AF_IUCV addressing mode, HiperSockets devices in layer 3 mode are identified through their hsuid sysfs attribute.

You set up a HiperSockets device for AF_IUCV by assigning a value to this attribute (see “Configuring a HiperSockets device for AF_IUCV addressing” on page 269).

### Setting up your z/VM guest virtual machine for IUCV

You must specify suitable IUCV statements for your z/VM guest virtual machine.

For details and for general IUCV setup information for z/VM guest virtual machines, see z/VM CP Programming Services, SC24-6179 and z/VM CP Planning and Administration, SC24-6178.

### Granting IUCV authorizations

Use the IUCV statement to grant the necessary authorizations.

**IUCV ALLOW**
allows any other z/VM virtual machine to establish a communication path with this z/VM virtual machine. With this statement, no further authorization is required in the z/VM virtual machine that initiates the communication.

**IUCV ANY**
allows this z/VM guest virtual machine to establish a communication path with any other z/VM guest virtual machine.

**IUCV <user ID>**
allows this z/VM guest virtual machine to establish a communication path to the z/VM guest virtual machine with the z/VM user ID <user ID>.

You can specify multiple IUCV statements. To any of these IUCV statements you can append the MSGLIMIT <limit> parameter. <limit> specifies the maximum number of outstanding messages that are allowed for each connection that is authorized by the statement. If no value is specified for MSGLIMIT, AF_IUCV requests 65 535, which is the maximum that is supported by IUCV.

### Setting a connection limit

Use the OPTION statement to limit the number of concurrent connections.

**OPTION MAXCONN <maxno>**

<maxno> specifies the maximum number of IUCV connections that are allowed for this virtual machine. The default is 64. The maximum is 65 535.
Example

These sample statements allow any z/VM guest virtual machine to connect to your z/VM guest virtual machine with a maximum of 10 000 outstanding messages for each incoming connection. Your z/VM guest virtual machine is permitted to connect to all other z/VM guest virtual machines. The total number of connections for your z/VM guest virtual machine cannot exceed 100.

IUCV ALLOW MSGLIMIT 10000
IUCV ANY
OPTION MAXCONN 100

Loading the IUCV modules

Red Hat Enterprise Linux 7.5 loads the af_iucv module when an application requests a socket in the AF_IUCV domain.

You can also use the `modprobe` command to load the AF_IUCV address family support module `af_iucv`:

```
# modprobe af_iucv
```

Addressing AF_IUCV sockets in applications

To use AF_IUCV sockets in applications, you must code a special AF_IUCV sockaddr structure.

**Application programmers:** This information is intended for programmers who want to use connections that are based on AF_IUCV addressing in their applications.

The primary difference between AF_IUCV sockets and TCP/IP sockets is how communication partners are identified (for example, how they are named). To use the AF_IUCV support in an application, code a sockaddr structure with AF_IUCV as the socket address family and with AF_IUCV address information.

For details, see the `af_iucv` man page.
Chapter 20. RDMA over Converged Ethernet

Linux on Z supports RDMA over Converged Ethernet (RoCE) in the form of 10GbE RoCE Express features.

A 10GbE RoCE Express feature physically consists of a Mellanox ConnectX-3 EN or Mellanox ConnectX-4 adapter. The adapters are two-port Ethernet adapters. On a mainframe, the mapping of ports to function keys depend on the adapter hardware:

- The two Mellanox ConnectX-3 EN adapter ports belong to the same function ID.
- The two Mellanox ConnectX-4 adapter ports belong to different function IDs.

The RoCE support requires PCI Express support, see "PCI Express support" on page 19.

Working with the RoCE support

Because the 10 GBE RoCE Express feature hardware physically consists of a Mellanox adapter, you must ensure that the following prerequisites are fulfilled before you can work with it.

Procedure

1. Ensure that PCIe support is enabled and the PCI card is active on your system. See “Setting up the PCIe support” on page 19 and “Using PCIe hotplug” on page 20.

2. Use the appropriate Mellanox device driver:
   - If you want to use TCP/IP, you need the mlx4_core and mlx4_en or mlx5_core module. If it is not already loaded, load it using for example, modprobe.
   - If you also want to use RDMA with InfiniBand (that is, using reliable datagram sockets, RDS), you need the mlx4_ib or mlx5_ib module. If it is not compiled into kernel or already loaded, load it using for example, modprobe. To use RDS, you also need the rds module and the rds_rdma module, see https://www.openfabrics.org/index.php/openfabrics-software.html.

3. Activate the network interface. You need to know the network interface name, which you can find under:

Use the ip command or equivalent to activate the interface. See the dev_port sysfs attribute of the interface name to ensure that you are working with the correct port.

What to do next

For further information about Mellanox, see:

Enabling debugging

The Mellanox mlx4 device driver can be configured with a kernel configuration option for debugging.

About this task

Debugging for the Mellanox mlx4 device driver is only available if the device driver is compiled with the kernel-configuration menu option CONFIG_MLX4_DEBUG.

Procedure

1. Check that the device driver has the CONFIG_MLX4_DEBUG option enabled.
2. Load the mlx4 modules with the sysfs parameter debug_level=1 to write debug messages to the syslog. Check the value of the debug_level parameter. If the parameter is set to 0, you can set it to 1 with the following command:
   
   ```bash
   echo 1 > /sys/module/mlx4_core/parameters/debug_level
   ```
## Part 5. System resources

| Chapter 21. Managing CPUs | 327 |
| Simultaneous multithreading | 327 |
| CPU capability change | 328 |
| Changing the configuration state of CPUs | 328 |
| Setting CPUs online or offline | 329 |
| Examining the CPU topology | 330 |
| CPU polarization | 331 |

| Chapter 22. Managing hotplug memory | 333 |
| What you should know about memory hotplug | 333 |
| Setting up hotplug memory | 334 |
| Performing memory management tasks | 334 |

| Chapter 23. Large page support | 339 |
| Setting up hugetlbfs large page support | 339 |
| Working with hugetlbfs large page support | 341 |

| Chapter 24. S/390 hypervisor file system | 343 |
| Directory structure | 343 |
| Setting up the S/390 hypervisor file system | 346 |
| Working with the S/390 hypervisor file system | 346 |

| Chapter 25. ETR- and STP-based clock synchronization | 349 |
| Enabling clock synchronization when booting | 349 |
| Enabling and disabling clock synchronization | 351 |

| Chapter 26. Identifying the IBM Z hardware | 353 |

| Chapter 27. HMC media device driver | 355 |
| Setting up the HMC media device driver | 355 |
| Working with the HMC media | 356 |

| Chapter 28. Data compression with GenWQE and zEDC Express | 359 |
| Features | 359 |
| What you should know about GenWQE | 359 |
| Setting up GenWQE hardware acceleration | 362 |
| Examples for using GenWQE | 363 |
| GenWQE hardware-acceleration for IBM Java | 365 |
| Exploring the GenWQE setup | 365 |
| External programming interfaces | 367 |

These device drivers and features help you to manage the resources of your real or virtual hardware.

### Newest version

You can find the newest version of this book at


### Restrictions

For prerequisites and restrictions see the IBM Z architecture specific information in the Red Hat Enterprise Linux 7.5 release notes at

Chapter 21. Managing CPUs

You can read CPU capability, activate standby CPUs, and examine the CPU topology.

Use the `lscpu` and `chcpu` commands to manage CPUs. These commands are part of the util-linux package. For details, see the man pages. Alternatively, you can manage CPUs through the attributes of their entries in sysfs.

Some attributes that govern CPUs are available in sysfs under:

```
/sys/devices/system/cpu/cpu<N>
```

where `<N>` is the number of the logical CPU. Both the sysfs interface and the `lscpu` and `chcpu` commands manage CPUs through their logical representation in Linux.

You can obtain a mapping of logical CPU numbers to physical CPU addresses by issuing the `lscpu` command with the `-e` option.

**Example:**

```
# lscpu -e
CPU NODE DRAWER BOOK SOCKET CORE L1d:L1i:L2d:L2i ONLINE CONFIGURED POLARIZATION ADDRESS
0 1 0 0 0 0 0:0:0:0 yes yes horizontal 0
1 1 0 0 0 0 1:1:1:1 yes yes horizontal 1
2 1 0 0 0 1 2:2:2:2 yes yes horizontal 2
3 1 0 0 0 1 3:3:3:3 yes yes horizontal 3
4 1 0 0 0 2 4:4:4:4 yes yes horizontal 4
5 1 0 0 0 2 5:5:5:5 yes yes horizontal 5
6 1 0 0 0 3 6:6:6:6 yes yes horizontal 6
7 1 0 0 0 3 7:7:7:7 yes yes horizontal 7
8 0 1 1 1 4 8:8:8:8 yes yes horizontal 8
... 
```

The logical CPU numbers are shown in the CPU column and the physical address in the ADDRESS column of the output table.

Alternatively, you can find the physical address of a CPU in the sysfs address attribute of a logical CPU.

**Example:**

```
# cat /sys/devices/system/cpu/cpu0/address
0
```

**Simultaneous multithreading**

Linux in LPAR mode can use the simultaneous multithreading technology on mainframes.

IBM z13 introduced the simultaneous multithreading technology to the mainframe. In Linux terminology, simultaneous multithreading is also known as SMT or Hyper-Threadiing.
With multithreading enabled, a single core on the hardware is mapped to multiple logical CPUs on Linux. Thus, multiple threads can issue instructions to a core simultaneously during each cycle.

To find out whether multithreading is enabled for a particular Linux instance, compare the number of cores with the number of threads that are available in the LPAR. You can use the `hyptop` command to obtain this information.

Simultaneous multithreading is designed to enhance performance. Whether this goal is achieved strongly depends on the available resources, the workload, and the applications that run on a particular Linux instance. Depending on these conditions, it might be advantageous to not make full use of multithreading or to disable it completely. Use the `hyptop` command to obtain utilization data for threads while Linux runs with multithreading enabled.

You can use the `smt=` and `nosmt` kernel parameters to control multithreading. By default, Linux in LPAR mode uses multithreading if it is provided by the hardware.

### CPU capability change

When the CPUs of a mainframe heat or cool, the Linux kernel generates a uevent for all affected online CPUs.

You can read the CPU capability from the Capability and, if present, Secondary Capability fields in `/proc/sysinfo`.

The capability value is an unsigned integer as defined in the system information block (SYSIB) 1.2.2 (see z/Architecture Principles of Operation, SA22-7832). A smaller value indicates a proportionally greater CPU capacity. Beyond that, there is no formal description of the algorithm that is used to generate this value. The value is used as an indication of the capability of the CPU relative to the capability of other CPU models.

### Changing the configuration state of CPUs

A CPU on an LPAR can be in a configured, standby, or reserved state. You can change the state of standby CPUs to configured state and vice versa.

**Before you begin**

- You can change the configuration state of CPUs for Linux in LPAR mode only. For Linux on z/VM, CPUs are always in a configured state.
- Daemon processes like `cpuplugd` can change the state of any CPU at any time. Such changes can interfere with manual changes.

**About this task**

When Linux is booted, only CPUs that are in a configured state are brought online and used. The kernel does not detect CPUs in reserved state.

**Procedure**

Issue a command of this form to change the configuration state of a CPU:

```
# chcpu [-c|-g] <N>
```
where

\(<N>\)

is the number of the logical CPU.

- `c` changes the configuration state of a CPU from standby to configured.

- `g` changes the configuration state of a CPU from configured to standby. Only offline CPUs can be changed to the standby state.

Alternatively, you can write 1 to the `configure` sysfs attribute of a CPU to set its configuration state to configured, or 0 to change its configuration state to standby.

Examples:

- The following `chcpu` command changes the state of the logical CPU with number 2 from standby to configured:

  ```
  # chcpu -c 2
  ```

The following command achieves the same results by writing 1 to the `configure` sysfs attribute of the CPU.

  ```
  # echo 1 > /sys/devices/system/cpu/cpu2/configure
  ```

- The following `chcpu` command changes the state of the logical CPU with number 2 from configured to standby:

  ```
  # chcpu -g 2
  ```

The following command achieves the same results by writing 0 to the `configure` sysfs attribute of the CPU.

  ```
  # echo 0 > /sys/devices/system/cpu/cpu2/configure
  ```

---

### Setting CPUs online or offline

Use the `chcpu` command or the `online` sysfs attribute of a logical CPU to set a CPU online or offline.

**Before you begin**

- Daemon processes like `cpuplugd` can change the state of any CPU at any time. Such changes can interfere with manual changes.

**Procedure**

1. Optional: Rescan the CPUs to ensure that Linux has a current list of configured CPUs.

   To initiate a rescan, issue the `chcpu` command with the `-r` option.

   ```
   # chcpu -r
   ```

   Alternatively, you can write 1 to `/sys/devices/system/cpu/rescan`.

   You might need a rescan for Linux on z/VM after one or more CPUs have been added to the z/VM guest virtual machine by the z/VM hypervisor. Linux in LPAR mode automatically detects newly available CPUs.
2. Change the online state of a CPU by issuing a command of this form:

```
# chcpu -e|-d <#/>
```

where

<#/>

is the number of the logical CPU.

- **-e** sets an offline CPU online. Only CPUs that are in the configuration state configured can be set online. For Linux on z/VM, all CPUs are in the configured state.

- **-d** sets an online CPU offline.

Alternatively, you can write 1 to the online sysfs attribute of a CPU to set it online, or 0 to set it offline.

**Examples:**

- The following **chcpu** commands force a CPU rescan, and then set the logical CPU with number 2 online.

```
# chcpu -r
# chcpu -e 2
```

The following commands achieve the same results by writing 1 to the online sysfs attribute of the CPU.

```
# echo 1 > /sys/devices/system/cpu/rescan
# echo 1 > /sys/devices/system/cpu/cpu2/online
```

- The following **chcpu** command sets the logical CPU with number 2 offline.

```
# chcpu -d 2
```

The following command achieves the same results by writing 0 to the online sysfs attribute of the CPU.

```
# echo 0 > /sys/devices/system/cpu/cpu2/online
```

---

**Examining the CPU topology**

If supported by your hardware, an interface is available that you can use to get information about the CPU topology of an LPAR.

**Before you begin**

Meaningful CPU topology information is available only to Linux in LPAR mode.

**About this task**

Use this information, for example, to optimize the Linux scheduler, which bases its decisions on which process gets scheduled to which CPU. Depending on the workload, this optimization might increase cache hits and therefore overall performance.
**Note**: By default CPU topology support is enabled in the Linux kernel. If it is not suitable for your workload, disable the support by specifying the kernel parameter `topology=off` in your parmfile or zipl.conf. See “Specifying kernel parameters” on page 23 for information about specifying kernel parameters.

The following sysfs attributes provide information about the CPU topology:

```plaintext
/sys/devices/system/cpu/cpu<N>/topology/thread_siblings
/sys/devices/system/cpu/cpu<N>/topology/core_siblings
/sys/devices/system/cpu/cpu<N>/topology/book_siblings
/sys/devices/system/cpu/cpu<N>/topology/drawer_siblings
```

where `<N>` specifies a particular logical CPU number. These attributes contain masks that specify sets of CPUs.

Because the mainframe hardware is evolving over time, the terms `drawer`, `book`, `core`, and `thread` do not necessarily correspond to fixed hardware entities. What matters for the Linux scheduler is the levels of relatedness that these terms signify, not the physical embodiment of the levels. In this context, more closely related means sharing more resources, like caches.

The `thread_siblings`, `core_siblings`, `book_siblings`, and `drawer_siblings` attribute each contain a mask that specifies the CPU and its peers at a particular level of relatedness.

1. The `thread_siblings` attribute covers the CPU and its closely related peers.
2. The `core_siblings` attribute covers all CPUs of the `thread_siblings` attribute and peers related at the core level.
3. The `book_siblings` attribute covers all CPUs of the `core_siblings` attribute and peers related at the book level.
4. The `drawer_siblings` attribute covers all CPUs of the `book_siblings` attribute and peers related at the drawer level.

If a machine reconfiguration causes the CPU topology to change, change uevents are created for each online CPU.

If the kernel also supports standby CPU activation and deactivation (see “Changing the configuration state of CPUs” on page 328), the masks also contain the CPUs that are in a configured, but offline state. Updating the masks after a reconfiguration might take some time.

---

**CPU polarization**

You can optimize the operation of a vertical SMP environment by adjusting the SMP factor based on the workload demands.

**Before you begin**

CPU polarization is relevant only to Linux in LPAR mode.

**About this task**

Horizontal CPU polarization means that the underlying hypervisor dispatches each virtual CPU of all z/VM guest virtual machines for the same amount of time.

If vertical CPU polarization is active, the hypervisor dispatches certain CPUs for a longer time than others for maximum performance. For example, if a guest has
three virtual CPUs, each of them with a share of 33%, then in case of vertical CPU polarization all of the processing time would be combined to a single CPU, which would run all the time, while the other two CPUs would get nearly no CPU time.

There are three types of vertical CPUs: high, medium, and low. Low CPUs hardly get any real CPU time, while high CPUs get a full real CPU. Medium CPUs get something in between.

**Note:** Running a system with different types of vertical CPUs can result in significant performance regressions. If possible, use only one type of vertical CPUs. Set all other CPUs offline and deconfigure them.

**Procedure**

To change the polarization, issue a command of this form:

```
# chcpu -p horizontal|vertical
```

Alternatively, you can write a 0 for horizontal polarization (the default) or a 1 for vertical polarization to `/sys/devices/system/cpu/dispatching`.

**Example:** The following `chcpu` command sets the polarization to `vertical`.

```
# chcpu -p vertical
```

You can achieve the same results by issuing the following command:

```
# echo 1 > /sys/devices/system/cpu/dispatching
```

**What to do next**

You can issue the `lscpu` command with the `-e` option to find out the polarization of your CPUs. For more detailed information for a particular CPU, read the `polarization` attribute of the CPU in sysfs.

```
# cat /sys/devices/system/cpu/cpu<N>/polarization
```

The polarization can have one of the following values:

- **horizontal** - each of the guests' virtual CPUs is dispatched for the same amount of time.
- **vertical:high** - full CPU time is allocated.
- **vertical:medium** - medium CPU time is allocated.
- **vertical:low** - very little CPU time is allocated.
- **unknown** - temporary value following a polarization change until the change is completed and the kernel has established the new polarization of each CPU.
Chapter 22. Managing hotplug memory

You can dynamically increase or decrease the memory for your running Linux instance.

To make memory available as hotplug memory, you must define it to your LPAR or z/VM. Hotplug memory is supported by z/VM 5.4 with the PTF for APAR VM64524 and by later z/VM versions.

For more information about memory hotplug, see Documentation/memory-hotplug.txt in the Linux source tree.

What you should know about memory hotplug

Hotplug memory is represented in sysfs. After rebooting Linux, all hotplug memory is offline.

Hotplug memory management overhead

Linux requires 64 bytes of memory to manage a 4-KB page of hotplug memory.

Use the following formula to calculate the total amount of initial memory that is consumed to manage your hotplug memory:

<hotplug_memory> / 64

Example: 4.5 TB of hotplug memory consume 4.5 TB / 64 = 72 GB.

For large amounts of hotplug memory, you might have to increase the initial memory that is available to your Linux instance. Otherwise, booting Linux might fail with a kernel panic and a message that there is not enough free memory.

How memory is represented in sysfs

Both the core memory of a Linux instance and the available hotplug memory are represented by directories in sysfs.

The memory with which Linux is started is the core memory. On the running Linux system, additional memory can be added as hotplug memory. The Linux kernel requires core memory to allocate its own data structures.

In sysfs, both the core memory of a Linux instance and the available hotplug memory are represented in form of memory blocks of equal size. Each block is represented as a directory of the form /sys/devices/system/memory/memory<n>, where <n> is an integer. You can find out the block size by reading the /sys/devices/system/memory/block_size_bytes attribute.

In the naming scheme, the memory blocks with the lowest address ranges are assigned the lowest integer numbers. The core memory always begins with memory0. The hotplug memory blocks follow the core memory blocks.

You can calculate where the hotplug memory begins. To find the number of core memory blocks, divide the base memory by the block size.

Example:
• With a core memory of 512 MB and a block size of 128 MB, the core memory is represented by four blocks, memory0 through memory3. Therefore, first hotplug memory block on this Linux instance is memory4.

• Another Linux instance with a core memory of 1024 MB and access to the same hotplug memory, represents this first hotplug memory block as memory8.

The hotplug memory is available to all operating system instances within the z/VM system or LPAR to which it was defined. The state sysfs attribute of a memory block indicates whether the block is in use by your own Linux system. The state attribute does not indicate whether a block is in use by another operating system instance. Attempts to add memory blocks that are already in use fail.

**Hotplug memory and reboot**

The original core memory is preserved as core memory and hotplug memory is freed when rebooting a Linux instance.

When you perform an IPL after shutting down Linux, always use `ipl clear` to preserve the original memory configuration.

---

**Setting up hotplug memory**

Before you can use hotplug memory on your Linux instance, you must define this memory as hotplug memory on your physical or virtual hardware.

**Defining hotplug memory to an LPAR**

You use the Hardware Management Console (HMC) to define hotplug memory as `reserved storage` on an LPAR.

For information about defining reserved storage for your LPAR, see the *Processor Resource/Systems Manager™ Planning Guide*, SB10-7041 for your mainframe.

**Defining hotplug memory to z/VM**

In z/VM, you define hotplug memory as `standby storage`.

There is also `reserved storage` in z/VM, but other than reserved memory defined for an LPAR, reserved storage that is defined in z/VM is not available as hotplug memory.

Always align the z/VM guest storage with the Linux memory block size. Otherwise, memory blocks might be missing or impossible to set offline in Linux.

For information about defining standby memory for z/VM guests, see the “DEFINE STORAGE” section in *z/VM CP Commands and Utilities Reference*, SC24-6175.

---

**Performing memory management tasks**

Typical memory management tasks include finding out the memory block size, adding memory, and removing memory.

- “Finding out the memory block size” on page 335
- “Listing the available memory blocks” on page 335
- “Adding memory” on page 336
Finding out the memory block size

On an IBM Z mainframe, memory is provided to Linux as memory blocks of equal size.

Procedure

- Use the `lsmem` command to find out the size of your memory blocks (see “lsmem: Show online status information about memory blocks” on page 579).

Example:

```
# lsmem
```

<table>
<thead>
<tr>
<th>Address range</th>
<th>Size (MB)</th>
<th>State</th>
<th>Removable</th>
<th>Device</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x0000000000000000-0x00000ffffff</td>
<td>256</td>
<td>online</td>
<td>no</td>
<td>0</td>
</tr>
<tr>
<td>0x0000000010000000-0x000000002fffffff</td>
<td>512</td>
<td>online</td>
<td>yes</td>
<td>1-2</td>
</tr>
<tr>
<td>0x0000000030000000-0x000000003fffffff</td>
<td>256</td>
<td>online</td>
<td>no</td>
<td>3</td>
</tr>
<tr>
<td>0x0000000040000000-0x000000005fffffff</td>
<td>768</td>
<td>online</td>
<td>yes</td>
<td>4-6</td>
</tr>
<tr>
<td>0x0000000070000000-0x000000008fffffff</td>
<td>2304</td>
<td>offline</td>
<td>-</td>
<td>7-15</td>
</tr>
</tbody>
</table>

Memory device size : 256 MB
Memory block size  : 256 MB
Total online memory : 1792 MB
Total offline memory: 2304 MB

In the example, the block size is 256 MB.

- Alternatively, you can read `/sys/devices/system/memory/block_size_bytes`. This sysfs attribute contains the block size in byte in hexadecimal notation.

Example:

```
# cat /sys/devices/system/memory/block_size_bytes
10000000
```

This hexadecimal value corresponds to 256 MB.

Listing the available memory blocks

List the available memory to find out how much memory is available and which memory blocks are online.

Procedure

- Use the `lsmem` command to list your memory blocks.

Example:
# lsmem -a

Address range | Size (MB) | State | Removable | Device
---|---|---|---|---
0x0000000000000000-0x000000000fffffff | 256 | online | no | 0
0x0000000010000000-0x000000001fffffff | 256 | online | no | 1
0x0000000020000000-0x000000002fffffff | 256 | online | no | 2
0x0000000030000000-0x000000003fffffff | 256 | online | yes | 3
0x0000000040000000-0x000000004fffffff | 256 | online | yes | 4
0x0000000050000000-0x000000005fffffff | 256 | offline | - | 5
0x0000000060000000-0x000000006fffffff | 256 | offline | - | 6
0x0000000070000000-0x000000007fffffff | 256 | offline | - | 7

Memory device size: 256 MB
Memory block size: 256 MB
Total online memory: 1280 MB
Total offline memory: 786 MB

For more information about the lsmem command, see "lsmem - Show online status information about memory blocks" on page 579.

- Alternatively, you can list the available memory blocks by listing the contents of /sys/devices/system/memory. Read the state attributes of each memory block to find out whether it is online or offline.

**Example:** The following command results in an overview for all available memory blocks.

```
# grep -r --include="state" "line" /sys/devices/system/memory/
```

**Note**

Online blocks are in use by your Linux instance. An offline block can be free to be added to your Linux instance but it might also be in use by another Linux instance.

**Adding memory**

You can add memory to your Linux instance by setting unused memory blocks online.

**Suspend and resume:**

Do not add hotplug memory if you intend to suspend the Linux instance before the next IPL. Any changes to the original memory configuration prevent suspension, even if you restore the original memory configuration by removing memory blocks that have been added. See Chapter 7, "Suspending and Resuming Linux," on page 103 for more information about suspending and resuming Linux.

**Procedure**

- Use the chmem command with the -e parameter to set memory online. You can specify the amount of memory you want to add with the command without specifying particular memory blocks. If there are enough eligible memory blocks
to satisfy your request, the tool finds them for you and sets the most suitable blocks online. For information about the \texttt{chmem} command, see "chmem - Set memory online or offline" on page 494.

- Alternatively, you can write \texttt{online} to the sysfs \texttt{state} attribute of an unused memory block. Issue a command of the form:

  \begin{verbatim}
  # echo online > /sys/devices/system/memory/memory<n>/state
  \end{verbatim}

  where \texttt{<n>} is an integer that identifies the memory unit.

\section*{Results}

Adding the memory block fails if the memory block is already in use. The \texttt{state} attribute changes to online when the memory block has been added successfully.

\section*{Removing memory}

You can remove memory from your Linux instance by setting memory blocks offline.

\section*{About this task}

Avoid removing core memory. The Linux kernel requires core memory to allocate its own data structures.

\section*{Procedure}

- Use the \texttt{chmem} command with the \texttt{-d} parameter to set memory offline. You can specify the amount of memory you want to remove with the command without specifying particular memory blocks. The tool finds eligible memory blocks for you and sets the most suitable blocks offline. For information about the \texttt{chmem} command, see "chmem - Set memory online or offline" on page 494.

- Alternatively, you can write \texttt{offline} to the sysfs \texttt{state} attribute of an unused memory block. Issue a command of the form:

  \begin{verbatim}
  # echo offline > /sys/devices/system/memory/memory<n>/state
  \end{verbatim}

  where \texttt{<n>} is an integer that identifies the memory unit.

\section*{Results}

The hotplug memory functions first relocate memory pages to free the memory block and then remove it. The \texttt{state} attribute changes to offline when the memory block has been removed successfully.

The memory block is not removed if it cannot be freed completely.
Chapter 23. Large page support

Large page support entails support for the Linux hugetlbfs file system.

The large page support virtual file system is backed by larger memory pages than the usual 4 K pages; for IBM Z the hardware page size is 1 MB.

To check whether 1 MB large pages are supported in your environment, issue the command:

```
# grep edat /proc/cpuinfo
```

features : esan3 zarch stfle msa ldisp eimm dfp edat etf3eh highgprs te

An output line that lists edat as a feature indicates 1 MB large page support.

Applications that use large page memory save a considerable amount of page table memory. Another benefit from the support might be an acceleration in the address translation and overall memory access speed.

As of version 7, Red Hat Enterprise Linux also supports libhugetlbfs linking. For more information, see the libhugetlbfs package, libhugetlbfs-

<version>.s390x.rpm, and the how-to document that is included in the package.

As of version 7, Red Hat Enterprise Linux supports transparent hugepages. For more information, see Documentation/vm/transhuge.txt in the Linux source tree.

As of zEC12, you can also configure 2 GB large pages if Linux is running on an LPAR. There is no flag that indicates 2GB support; the support is always there as of zEC12. See "Pre-allocating 2 GB large pages" on page 340.

Setting up hugetlbfs large page support

You configure hugetlbfs large page support by adding parameters to the kernel parameter line.

```
Large page support kernel parameter syntax

default_hugepagesz=<size>

hugepages=<number>
```

where:

**default_hugepagesz=<size>**

specifies the default page size in byte. You can use suffixes K, M, and G to specify KB, MB, and GB. The default value is 1 MB.

**hugepages=<number>**

is the number of large pages to be allocated at boot time.
hugepagesz=<size>
    specifies the page size in byte. You can use suffixes K, M, and G to specify KB, MB, and GB.

Note: If you specify more pages than available, Linux reserves as many as possible. As a likely result, too few general pages remain for the boot process, and your system stops with an out-of-memory error.

Pre-allocating 2 GB large pages

If Linux is running in an LPAR, you can use 2 GB large pages.

Before you can use 2 GB large pages, you must pre-allocate them to the kernel page pool. To pre-allocate 2 GB pages, precede the hugepages= parameter with the page size selection parameter, hugepagesz=2G.

Tip: Memory quickly becomes fragmented after booting, and new 2 GB large pages cannot be allocated. Use kernel boot parameters to allocate 2 GB large pages to avoid the memory fragmentation problem.

To pre-allocate a number of pages of 2 GB size and also set the default size to 2 GB:

    default_hugepagesz=2G hugepagesz=2G hugepages=<number>

Setting up multiple large page pools

You can allocate multiple large page pools and use them simultaneously. To allocate multiple large page pools, specify the hugepagesz= parameter several times, each time followed by a corresponding hugepages= parameter.

For example, to specify two pools, one with 1 MB pages and one with 2 GB pages, specify:

    hugepagesz=1M hugepages=8 hugepagesz=2G hugepages=2

This creates a sysfs attribute for each pool, /sys/kernel/mm/hugepages/hugepages-
<size>kB/nr_hugepages, where <size> is the page size in KB. For the example given, the following attributes are created:

    /sys/kernel/mm/hugepages/hugepages-1024kB/nr_hugepages
    /sys/kernel/mm/hugepages/hugepages-2097152kB/nr_hugepages

Large pages and hotplug memory

Hotplug memory that is added to a running Linux instance is movable and can be allocated to movable resources only.

By default, large pages are not movable and cannot be allocated from movable memory. You can enable allocation from movable memory with the sysctl setting hugepages_treat_as_movable.

To enable allocation of large pages from movable hotplug memory, issue:

    # echo 1 > /proc/sys/vm/hugepages_treat_as_movable
Although this setting makes large pages eligible for allocation through movable memory, it does not make large pages movable. As a result, the allocated hotplug memory cannot be set offline until all large pages are released from that memory.

To disable allocation of large pages from movable hotplug memory, issue:

```
# echo 0 > /proc/sys/vm/hugepages_treat_as_movable
```

---

**Working with hugetlbfs large page support**

Typical tasks for working with hugetlbfs large page support include reading the current number of large pages, changing the number of large pages, and display information about available large pages.

**About this task**

The large page memory can be used through `mmap()` or SysV shared memory system calls, more detailed information can be found in the Linux kernel source tree under `Documentation/vm/hugetlbpage.txt`, including implementation examples.

Your database product might support large page memory. See your database documentation to find out if and how it can be configured to use large page memory.

Depending on your version of Java™, you might require specific options to make a Java program use the large page feature. For IBM SDK, Java Technology Edition 7, specify the `-Xlp` option. If you use the SysV shared memory interface, which includes `java -Xlp`, you must adjust the shared memory allocation limits to match the workload requirements. Use the following `sysctl` attributes:

**/proc/sys/kernel/shmall**

Defines the global maximum amount of shared memory for all processes, specified in number of 4 KB pages.

**/proc/sys/kernel/shmmax**

Defines the maximum amount of shared memory per process, specified in number of bytes.

For example, the following commands would set both limits to 20 GB:

```
# echo 5242880 > /proc/sys/kernel/shmall
# echo 21474836480 > /proc/sys/kernel/shmmax
```

**Procedure**

- Specify the `hugepages=` kernel parameter with the number of large pages to be allocated at boot time. To read the current number of default large pages, issue:

```
# cat /proc/sys/vm/nr_hugepages
```

- To change the number of default-sized large pages dynamically during runtime, write to procfs:

```
# echo 20 > /proc/sys/vm/nr_hugepages
```
If there is not enough contiguous memory available to fulfill the request, the maximum number of large pages are reserved.

- To obtain information about the number of large pages currently available and the large page size, issue:

```bash
# cat /proc/meminfo
...
HugePages_Total: 20
HugePages_Free: 14
HugePages_Rsvd: 0
HugePages_Surp: 0
...
Hugepagesize: 1024 KB
...
```

- To adjust characteristics of a large-page pool, when more than one pool exists, use the sysfs attributes of the pool. These can be found under
  `/sys/kernel/mm/hugepages/hugepages-<size>/nr_hugepages`

  Where `<size>` is the page size in KB.

**Example**

To allocate 2 GB large pages:

1. Specify 2 GB large pages and pre-allocate them to the page pool at boot time. Use the following kernel boot parameters:

   ```
   default_hugepagesz=2G hugepagesz=2G hugepages=4
   ```

2. After booting, read `/proc/meminfo` to see information about the amount of large pages currently available and the large page size:

   ```bash
   cat /proc/meminfo
   ...
   HugePages_Total: 4
   HugePages_Free: 4
   HugePages_Rsvd: 0
   HugePages_Surp: 0
   Hugepagesize: 2097152 KB
   ...
Chapter 24. S/390 hypervisor file system

The S/390® hypervisor file system (hypfs) provides a mechanism to access LPAR and z/VM hypervisor data.

Directory structure

When the hypfs file system is mounted, the accounting information is retrieved and a file system tree is created. The tree contains a full set of attribute files with the hypervisor information.

By convention, the mount point for the hypervisor file system is /sys/hypervisor/s390.

LPAR directories and attributes

There are hypfs directories and attributes with hypervisor information for Linux in LPAR mode.

Figure 69 illustrates the file system tree that is created for LPAR.

Figure 69. The hypervisor file system for LPAR

update  Write-only file to trigger an update of all attributes.

cpus/ Directory for all physical cores.

cpus/<core_ID> Directory for one physical core. <core_ID> is the logical (decimal) core number.
**type**  Type name of physical core, such as CP or IFL.

**mgmtime**  
Physical-LPAR-management time in microseconds (LPAR overhead).

**hyp/**  Directory for hypervisor information.

**hyp/type**  
Type of hypervisor (LPAR hypervisor).

**systems/**  
Directory for all LPARs.

**systems/</lpar name>/**  
Directory for one LPAR.

**systems/</lpar name>/cpus/</core_ID>/**  
Directory for the virtual cores for one LPAR. The </core_ID> is the logical (decimal) core number.

**type**  Type of the logical core, such as CP or IFL.

**mgmtime**  
LPAR-management time. Accumulated number of microseconds during which a physical core was assigned to the logical core and the core time was consumed by the hypervisor and was not provided to the LPAR (LPAR overhead).

**cputime**  
Accumulated number of microseconds during which a physical core was assigned to the logical core and the core time was consumed by the LPAR.

**onlinetime**  
Accumulated number of microseconds during which the logical core has been online.

**Note:** For LPARs with multithreading enabled, the entities in the cpus directories represent hardware cores, not threads.

**Note:** For older machines, the onlinetime attribute might be missing. Generally, it is advantageous for applications to tolerate missing attributes or new attributes that are added to the file system. To check the content of the files, you can use tools such as **cat** or **less**.

**z/VM directories and attributes**

There are hypfs directories and attributes with hypervisor information for Linux on z/VM.

**update**  Write-only file to trigger an update of all attributes.

**cpus/**  Directory for all physical CPUs.

**cpus/count**  
Total current CPUs.

**hyp/**  Directory for hypervisor information.

**hyp/type**  
Type of hypervisor (z/VM hypervisor).
systems/
    Directory for all z/VM guest virtual machines.

systems/<guest name>/
    Directory for one z/VM guest virtual machine.

systems/<guest name>/onlinetime_us
    Time in microseconds that the guest virtual machine has been logged on.

systems/<guest name>/cpus/
    Directory for the virtual CPUs for one guest virtual machine.

capped
    Flag that shows whether CPU capping is on for the guest virtual machine (0 = off, 1 = soft, 2 = hard).

count
    Total current virtual CPUs in the guest virtual machine.

cputime_us
    Number of microseconds where the guest virtual machine CPU was running on a physical CPU.

dedicated
    Flag that shows if the guest virtual machine has at least one dedicated CPU (0 = no, 1 = yes).

weight_cur
    Current share of guest virtual machine (1-10000); 0 for ABSOLUTE SHARE guests.

weight_max
    Maximum share of guest virtual machine (1-10000); 0 for ABSOLUTE SHARE guests.

weight_min
    Number of operating CPUs. Do not be confused by the attribute name, which suggests a different meaning.

systems/<guest name>/samples/
    Directory for sample information for one guest virtual machine.

cpu_delay
    Number of CPU delay samples that are attributed to the guest virtual machine.

cpu_using
    Number of CPU using samples attributed to the guest virtual machine.

idle
    Number of idle samples attributed to the guest virtual machine.

mem_delay
    Number of memory delay samples that are attributed to the guest virtual machine.

other
    Number of other samples attributed to the guest virtual machine.

total
    Number of total samples attributed to the guest virtual machine.

systems/<guest name>/mem/
    Directory for memory information for one guest virtual machine.

max_KiB
    Maximum memory in KiB (1024 bytes).

min_KiB
    Minimum memory in KiB (1024 bytes).
**Setting up the S/390 hypervisor file system**

To use the file system, it must be mounted. You can mount the file system with the mount command or with an entry in `/etc/fstab`.

To mount the file system manually, issue the following command:

```bash
# mount none -t s390_hypfs <mount point>
```

where `<mount point>` is where you want the file system mounted. Preferably, use `/sys/hypervisor/s390`.

To mount hypfs by using `/etc/fstab`, add the following line:

```plaintext
none <mount point> s390_hypfs defaults 0 0
```

If your z/VM system does not support DIAG 2fc, the s390_hypfs is not activated and it is not possible to mount the file system. Instead, an error message like this is issued:

```
mount: unknown filesystem type 's390_hypfs'
```

To get data for all z/VM guests, privilege class B is required for the guest where hypfs is mounted. For non-class B guests, data is provided only for the local guest.

To get data for all LPARs, select the Global performance data control check box in the HMC or SE security menu of the LPAR activation profile. Otherwise, data is provided only for the local LPAR.

**Working with the S/390 hypervisor file system**

Typical tasks that you must perform when working with the S/390 hypervisor file system include defining access permissions and updating hypfs information.

- “Defining access permissions”
- “Updating hypfs information” on page 347

**Defining access permissions**

The root user usually has access to the hypfs file system. It is possible to explicitly define access permissions.

**About this task**

If no mount options are specified, the files and directories of the file system get the uid and gid of the user who mounted the file system (usually root). It is possible to explicitly define uid and gid by using the mount options `uid=<number>` and `gid=<number>`. 
Example

You can define uid=1000 and gid=2000 with the following mount command:

```
# mount none -t s390_hypfs -o "uid=1000,gid=2000" <mount point>
```

Alternatively, you can add the following line to the /etc/fstab file:

```
none <mount point> s390_hypfs uid=1000,gid=2000 0 0
```

The first mount defines uid and gid. Subsequent mounts automatically have the same uid and gid setting as the first one.

The permissions for directories and files are as follows:

- Update file: 0220 (--w--w----)
- Regular files: 0440 (-r--r-----)
- Directories: 0550 (dr-xr-x---)

Updating hypfs information

You trigger the update process by writing something into the update file at the top-level hypfs directory.

Procedure

With hypfs mounted at /sys/hypervisor/s390, you can trigger the update process by issuing the following command:

```
# echo 1 > /sys/hypervisor/s390/update
```

During the update, the entire directory structure is deleted and rebuilt. If a file was open before the update, subsequent reads return the old data until the file is opened again. Within 1 second only one update can be done If multiple updates are triggered within a second, only the first update is performed and subsequent write system calls return -1 and errno is set to EBUSY. Applications can use the following procedure to ensure consistent data:

1. Read modification time through stat(2) from the update attribute.
2. If data is too old, write to the update attribute and start again with step 1.
3. Read data from file system.
4. Read modification time of the update attribute again and compare it with first timestamp. If the timestamps do not match, return to step 2.
Chapter 25. ETR- and STP-based clock synchronization

Your Linux instance might be part of an extended remote copy (XRC) setup that requires synchronization of the Linux time-of-day (TOD) clock with a timing network.

Linux on Z supports external time reference (ETR) and system time protocol (STP) based TOD synchronization. ETR and STP work independently of one another. If both ETR and STP are enabled, Linux might use either to synchronize the clock.

For more information about ETR, see the IBM Redbooks® technote at www.ibm.com/redbooks/abstracts/tips0217.html

For information about STP, see www.ibm.com/systems/z/advantages/pso/stp.html

Both ETR and STP support are included in Red Hat Enterprise Linux.

ETR requires at least one ETR unit that is connected to an external time source. For availability reasons, many installations use a second ETR unit. The ETR units correspond to two ETR ports on Linux. Always set both ports online if two ETR units are available.

Attention: Be sure that a reliable timing signal is available before enabling clock synchronization. With enabled clock synchronization, Linux expects regular timing signals and might stop indefinitely to wait for such signals if it does not receive them.

Enabling clock synchronization when booting

Use kernel parameters to enable clock synchronization when booting.

This section describes the kernel parameters that you can use to set up synchronization for your Linux TOD clock. These kernel parameters specify the initial synchronization settings. On a running Linux instance, you can change these settings through attributes in sysfs (see “Enabling and disabling clock synchronization” on page 351).
Enabling ETR-based clock synchronization

Use the `etr=` kernel parameter to set ETR ports online when Linux is booted.

ETR-based clock synchronization is enabled if at least one ETR port is online.

### etr syntax

- `etr=off`  
- `etr=on`  
- `etr=port0`  
- `etr=port1`

The values have the following effect:

- **on** sets both ports online.
- **port0**  
  sets `port0` online and `port1` offline.
- **port1**  
  sets `port1` online and `port0` offline.
- **off**  
  sets both ports offline. With both ports offline, ETR-based clock synchronization is not enabled. This is the default.

### Example

To enable ETR-based clock synchronization with both ETR ports online, specify:

```
etr=on
```
Enabling STP-based clock synchronization

Use the stp= kernel parameter to enable STP-based clock synchronization when Linux is booted.

**stp syntax**

- stp=off
- stp=on

By default, STP-based clock synchronization is not enabled.

**Example**

To enable STP-based clock synchronization, specify:

stp=on

---

Enabling and disabling clock synchronization

You can use the sysfs interfaces of ETR and STP to enable and disable clock synchronization on a running Linux instance.

### Enabling and disabling ETR-based clock synchronization

Use the ETR sysfs attribute `online` to set an ETR port online or offline.

**About this task**

ETR-based clock synchronization is switched off if both ETR ports are offline.

**Procedure**

To set an ETR port online, set its sysfs `online` attribute to 1. To set an ETR port offline, set its sysfs `online` attribute to 0. Enter a command of this form:

```
# echo <flag> > /sys/devices/system/etr/etr<n>/online
```

where `<n>` identifies the port and is either 0 or 1.

**Example**

To set ETR port `etr1` offline, enter:

```
# echo 0 > /sys/devices/system/etr/etr1/online
```

---

**Enabling and disabling STP-based clock synchronization**

Use the STP sysfs attribute `online` to enable or disable STP-based clock synchronization.
Procedure

To enable STP-based clock synchronization, set /sys/devices/system/stp/online to 1. To disable STP-based clock synchronization, set this attribute to 0.

Example

To disable STP-based clock synchronization, enter:

```
# echo 0 > /sys/devices/system/stp/online
```
In installations with several IBM Z mainframes, you might need to identify the particular hardware system on which a Linux instance is running.

Two attributes in /sys/firmware/ocf can help you to identify the hardware.

**cpc_name**
- contains the name that is assigned to the central processor complex (CPC).
- This name identifies the mainframe system on a Hardware Management Console (HMC).

**hmc_network**
- contains the name of the HMC network to which the mainframe system is connected.

The two attributes contain the empty string if the Linux instance runs as a guest of a hypervisor that does not support the operations command facility (OCF) communication parameters interface.

Use the **cat** command to read these attributes.

**Example:**

```
# cat /sys/firmware/ocf/cpc_name
Z05
# cat /sys/firmware/ocf/hmc_network
SNA00
```
Chapter 27. HMC media device driver

You use the HMC media device driver to access files on removable media at a system that runs the Hardware Management Console (HMC).

Before you begin: You must log in to the HMC on the system with the removable media and assign the media to the LPAR.

The HMC media device driver supports the following removable media:
- A DVD in the DVD drive of the HMC system
- A CD in the DVD drive of the HMC system
- USB-attached storage that is plugged into the HMC system

The most commonly used removable media at the HMC is a DVD.

The HMC media device driver uses the /dev/hmcdrv device node to support these capabilities:
- List the media contents with the lshmc command (see "lshmc - List media contents in the HMC media drive" on page 575).
- Mount the media contents as a file system with the hmcdrvfs command (see "hmcdrvfs - Mount a FUSE file system for remote access to media in the HMC media drive" on page 547).

Installers on suitably prepared installation DVDs can use these capabilities to install Linux in an LPAR.

Setting up the HMC media device driver

You can set the cache size for the HMC media device driver.

Kernel parameters

If the HMC media device driver has been compiled into the kernel, you set the cache size by adding the hmcdrv.cachesize= parameter to the kernel parameter line.

```
hmcdrv kernel parameter syntax

hmcdrv.cachesize=534288

hmcdrv.cachesize=<size>
```

where <size> is the cache size in bytes. The specification must be a multiple of 2048. You can use the suffixes K for kilobytes, M for megabytes, or G for gigabytes. Specify 0 to not cache any media content. By default, the cache size is 534288 bytes (0.5 MB).

Loading the hmcdrv module creates a device node at /dev/hmcdrv.
Module parameters

If the HMC media device driver has been built as a separate module, hmcdrv, you set the cache size through the cachesize= module parameter.

Before you can work with the HMC media device driver and with the dependent lshmc and hmcdrvfs commands, you must load the hmcdrv kernel module.

```
modprobe hmcdrv
```

**Example**

The following specifications are equivalent:

```
# modprobe hmcdrv cachesize=153600
# modprobe hmcdrv cachesize=150K
```

```hmcdrv module parameter syntax```

```modprobe hmcdrv cachesize=534288```

```cachesize=<size>```

```modprobe hmcdrv cachesize=153600```

```cachesize=150K```

where `<size>` is the cache size in bytes. The specification must be a multiple of 2048. You can use the suffixes K for kilobytes, M for megabytes, or G for gigabytes. Specify 0 to not cache any media content. By default, the cache size is 534288 bytes (0.5 MB).

Loading the hmcdrv module creates a device node at `/dev/hmcdrv`.

**Example**

The following specifications are equivalent:

```
# modprobe hmcdrv cachesize=153600
# modprobe hmcdrv cachesize=150K
```

**Working with the HMC media**

You can list files on media that are inserted into the HMC system and you can mount the media content on the Linux file system.

- “Assigning the removable media of the HMC to an LPAR”
- “Listing files on the removable media at the HMC” on page 357
- “Mounting the content of the removable media at the HMC” on page 358

**Assigning the removable media of the HMC to an LPAR**

Use the HMC to assign the removable media to the LPAR where your Linux instance runs.
Before you begin

- You need access to the HMC, and you must be authorized to use the Access Removable Media task for the LPAR to which you want to assign the media.
- For Linux on z/VM, the z/VM guest virtual machine must have at least privilege class B.
- For Linux in LPAR mode, the LPAR activation profile must allow issuing SCLP requests.

About this task

You can list files on the removable media at the HMC without having to first mount the contents on the Linux file system.

Procedure

1. Insert the removable media into the HMC system.
2. Use the Access Removable Media task on your HMC to assign the removable media to the LPAR where your Linux instance runs.
   
   For Linux on z/VM, this is the LPAR where the z/VM hypervisor runs that provides the guest virtual machine to your Linux instance.
   
   For details, see the HMC documentation for the HMC at your installation.

Results

You can now access the removable media from your Linux instance.

Listing files on the removable media at the HMC

Use the lshmc command to list files on the removable media at the HMC.

Before you begin

Your Linux instance must have access to the removable media at the HMC (see “Assigning the removable media of the HMC to an LPAR” on page 356).

About this task

You can list files on the removable media at the HMC without having to first mount the contents on the Linux file system.

Procedure

Issue a command of this form:

```
# lshmc <filepath>
```

where <filepath> is an optional specification for a particular path and file. Path specifications are interpreted as relative to the root directory of the removable media. You can use the asterisk (*) and question mark (?) as wildcards. If you omit <filepath>, all files in the root directory of the media are listed.

Example: The following command lists all .html files in the www subdirectory of the media.
# lshmc www/*.html

For more information about the `lshmc` command, see “lshmc - List media contents in the HMC media drive” on page 575.

Mounting the content of the removable media at the HMC

Use the `hmcdrvfs` command to mount the content of the removable media at the HMC.

**Before you begin**

Your Linux instance must have access to the removable media of the HMC (see “Assigning the removable media of the HMC to an LPAR” on page 356).

**About this task**

You can mount the content of the removable media at the HMC in read-only mode on the Linux file system.

**Procedure**

1. Optional: Confirm that you are accessing the intended content by issuing the `lshmc` command.
2. Mount the media content by issuing a command of this form:

   ```bash
   # hmcdrvfs <mountpoint>
   ```

   where `<mountpoint>` is the mount point on the Linux file system.

   **Example:** The following command mounts the media content at `/mnt/hmc`:

   ```bash
   # hmcdrvfs /mnt/hmc
   ```

**Results**

You can now access the files on the media in read-only mode through the Linux file system.

**What to do next**

When you no longer need access to the media content, unmount the media with the `fusermount` command.
Chapter 28. Data compression with GenWQE and zEDC Express

Generic Work Queue Engine (GenWQE) supports hardware-accelerated data compression and decompression through zEDC Express, a PCIe-attached Field Programmable Gate Array (FPGA) acceleration adapter.

zEDC Express is available for zEC12 and later IBM mainframes.

zEDC hardware-acceleration is available for both Linux and z/OS. For more information about zEDC on z/OS and about setting up zEDC Express, see Reduce Storage Occupancy and Increase Operations Efficiency with IBM zEnterprise Data Compression, SG24-8259. You can obtain this publication from the IBM Redbooks website at [www.redbooks.ibm.com/abstracts/sg248259.html](http://www.redbooks.ibm.com/abstracts/sg248259.html).

Features

GenWQE supports hardware-accelerated data compression and decompression with common standards.

- GenWQE implements the zlib API.
- GenWQE adheres to the following RFCs:
  - RFC 1950 (zlib)
  - RFC 1951 (deflate)
  - RFC 1952 (gzip)

These standards ensure compatibility among different zlib implementations.

- Data that is compressed with GenWQE can be decompressed through a zlib software library.
- Data that is compressed through a software zlib software library can be decompressed with GenWQE.

- GenWQE supports the following PCIe FPGA acceleration hardware:
  - zEDC Express

What you should know about GenWQE

Learn about the GenWQE components, how to enable GenWQE accelerated zlib for user applications, and device representation in Linux.

The GenWQE accelerated zlib

The GenWQE accelerated zlib can replace a zlib software library.

For data compression and decompression tasks, Red Hat Enterprise Linux 7.5 includes software libraries. The zlib library, which provides the zlib API, is one of the most commonly used libraries for data compression and decompression. For information about zlib, see [www.zlib.net](http://www.zlib.net).

Because the GenWQE accelerated zlib offers the zlib API, applications can use it instead of the default zlib software library. The GenWQE hardware-accelerated zlib is designed to enhance performance by offloading tasks to a hardware accelerator.
Applications

You can make the user space components of the GenWQE hardware-accelerated zlib available to applications that request data compression functions through the zlib API. Red Hat Enterprise Linux 7.5 provides these user space components with the genwqe-zlib RPM.

A second RPM, genwqe-tools, provides tools that use the GenWQE hardware-accelerated zlib.

IBM Java version 7.1 or later includes components of the GenWQE hardware-accelerated zlib. Through these components, it can directly address the GenWQE device nodes. With the required environment variables in place, it uses hardware-acceleration if it is available (see “GenWQE hardware-acceleration for IBM Java” on page 365).

Hardware-accelerated zlib

The hardware-accelerated zlib is a zlib implementation that acts as a wrapper for two included libraries:

libzHW

- a hardware library that prepares requests for processing by the hardware accelerator. The hardware library is intended to handle the bulk of the requests.
- This library also manages data buffers for optimized hardware compression.
libz — a software implementation of the zlib interface. Because it provides the same interface as its wrapper library, it can handle any requests unmodified.

The hardware-accelerated zlib arbitrates between the two included libraries. It uses the software library as a backup if no hardware accelerator is available. It also evaluates the expected performance gain against the extra processing for channeling requests to the accelerator. For small or fragmented data, software processing might be advantageous, especially for decompression. The evaluation takes available resources, such as buffer space, into account.

Card library

The card library, libcard, mediates between the hardware-accelerated zlib library and the GenWQE device driver. It provides recovery features and can move jobs between available accelerators.

Device driver

The GenWQE device driver is the kernel part of GenWQE. It serializes requests to an accelerator in form of device driver control blocks (DDCBs), and it enables multi-process and multi-thread usage.

GenWQE device nodes

GenWQE user space components use device nodes to exchange data with the GenWQE device driver.

Red Hat Enterprise Linux 7.5 automatically loads the GenWQE device driver module when it is required. It also creates a device node of the form /dev/genwqe<n>_card for each available virtual acceleration card. <n> is an index number that identifies an individual virtual card. Node /dev/genwqe0_card is assigned to the first card that is detected, /dev/genwqe1_card to the second card, and so on.

Do not directly use these device nodes. The nodes are intended to be used by the user space components of the GenWQE hardware-accelerated zlib and by IBM Java.

Virtual accelerators

Each physical accelerator card can provide up to 15 virtual cards. In PCIe terminology, these virtual cards are called virtual functions.

GenWQE accelerator cards, as detected by Linux on Z, are virtual cards. Which and how many cards are available to a particular Linux instance depends on the mainframe configuration and, if applicable, the hypervisor configuration.

As for most mainframe devices, availability can be enhanced by assigning virtual accelerator cards from different physical cards.

A degree of load distribution can be achieved by unevenly distributing accelerator cards among different Linux instances.

Tradeoff between best compression and speed

A minimum size of compressed data and fast compression are conflicting goals.
For hardware-accelerated compression with GenWQE, the compression ratio is roughly equivalent to `gzip --fast`.

Data that was compressed with GenWQE hardware-acceleration might have a different size from data that was compressed in software. The data compression standards are not violated by this difference. Despite possible differences in size of the compressed data, data that is compressed with GenWQE hardware-acceleration can be decompressed in software and vice versa.

**Setting up GenWQE hardware acceleration**

Install the GenWQE components and understand how environment variables can override default settings.

**Installing the GenWQE hardware-accelerated zlib**

Install the `genwqe-zlib` and `genwqe-tools` RPMs that are included in Red Hat Enterprise Linux 7.5.

The `genwqe-zlib` RPM includes the user space components of the GenWQE hardware-accelerated zlib.

The `genwqe-tools` RPM provides the following tools:

- `genwqe_gzip` and `genwqe_gunzip`, which are GenWQE versions of `gzip` and `gunzip` (see “Examples for using GenWQE” on page 363).

  These tools can be used for most purposes, but they do not implement all of the more unusual options of their common code counterparts. See the man pages to find out which options are supported.

- `genwqe_echo`, a tool to confirm the availability of accelerator hardware through the GenWQE accelerated zlib. See “Confirming that the accelerator hardware can be reached” on page 366 for details.

**Environment variables**

You can set environment variables to control the GenWQE hardware-accelerated zlib.

The GenWQE hardware-accelerated zlib uses defaults that correspond to the following environment variable settings:

- `ZLIB_ACCELERATOR=GENWQE`
- `ZLIB_CARD=-1`
- `ZLIB_TRACE=0x0`
- `ZLIB_DEFLATE_IMPL=0x41`
- `ZLIB_INFLATE_IMPL=0x41`

You can override these defaults by setting the following environment variables:

- `ZLIB_ACCELERATOR`  
  Sets the accelerator type. For zEDC Express, the type is `GENWQE`.

- `ZLIB_CARD`  
  -1, uses all accelerators that are available to the Linux instance. Failed requests are retried on alternative accelerators.

  You can specify the ID of a particular virtual accelerator card to be used. The ID is the index number that makes the nodes unique. All other cards are ignored, and no retry on alternative cards is performed if the specified card fails. Specify an ID only if you want to test a particular card.
uses the first card that is found by the device driver. As for specifying an individual card, all other cards are ignored.

**ZLIB_TRACE**
Sets tracing bits:
- \( 0x1 \) General trace.
- \( 0x2 \) Hardware trace.
- \( 0x4 \) Software trace.
- \( 0x8 \) Trace summary at the end of a process.

Tracing requires extra processing and incurs a performance penalty. The least performance impact is to be expected from the trace summary. By default, tracing is off.

**ZLIB_DEFLATE_IMPL**
\( 0x01 \) and \( 0x41 \) enable hardware compression, where \( 0x41 \) adds an optimization setting. \( 0x00 \) forces software compression and is intended for experimentation, for example, for gathering performance data with and without hardware acceleration.

**ZLIB_INFLATE_IMPL**
\( 0x01 \) and \( 0x41 \) enable hardware decompression, where \( 0x41 \) adds an optimization setting. \( 0x00 \) forces software decompression and is intended for experimentation, for example, for gathering performance data with and without hardware acceleration.

You can find more details about the environment variables in the GenWQE wiki on GitHub at [github.com/ibm-genwqe/genwqe-user/wiki/Environment Variables](https://github.com/ibm-genwqe/genwqe-user/wiki/Environment Variables).

### Examples for using GenWQE

You can use the GenWQE hardware-accelerated zlib through GenWQE tools.

**Activating the GenWQE hardware-accelerated zlib for an application**

Whether and how you can make an application use the GenWQE hardware-accelerated zlib depends on how the application links to \( \text{libz.so} \).

Examine the application for links to \( \text{libz.so} \), for example with the `ldd` tool.

- If the application does not link to \( \text{libz.so} \) or if it statically links to \( \text{libz.so} \), it would require reccompilation, and possibly code changes, to make acceleration through GenWQE possible.
- If an application dynamically links to \( \text{libz.so} \), you might be able to redirect the library calls from the default implementation to the GenWQE hardware-accelerated zlib.

Some applications require zlib features that are not available from the GenWQE hardware-accelerated zlib. Such applications fail if a global redirect is put in place. The following technique redirects calls for the scope of a particular application.

Specify the `LD_PRELOAD` environment variable to load the GenWQE hardware-accelerated zlib. Set the variable with the start command for your application.

**Example:**
Compressing data with genwqe_gzip

GenWQE provides two tools, genwqe_gzip and genwqe_gunzip that can be used in place of the common code gzip and gunzip tools. The GenWQE versions of the tools use hardware acceleration if it is available.

Procedure

Run the genwqe_gzip command with the -AGENWQE parameter to compress a file.

```
# genwqe_gzip -AGENWQE <file>
```

The -AGENWQE parameter ensures that the correct, PCIe-attached, accelerator card is used. Also use this option when decompressing data with the genwqe_gunzip command. See the man pages for other options.

Running tar with GenWQE hardware-acceleration

You can make tar use genwqe_gzip in place of the common code gzip.

About this task

If called with the z option, the tar utility uses the first gzip tool in the search path, which is usually the common code version. By inserting the path to the GenWQE gzip tool at the beginning of the PATH variable, you can make the tar utility use hardware acceleration.

The path points to /usr/lib64/genwqe/gzip and /usr/lib64/genwqe/gunzip, which are symbolic links to genwqe_gzip and genwqe_gunzip.

The acceleration is most marked for a single large text file. The example that follows compresses a directory with the Linux source code.

Procedure

1. Run the tar command as usual to use software compression. To obtain performance data, specify the tar command as an argument to the time command.

```
# time tar cfz linux-src.sw.tar.gz linux-src
real 0m22.329s
user 0m22.147s
sys 0m0.849s
```

2. Run the tar command with an adjusted PATH variable to use GenWQE hardware acceleration. Again, use the time command to obtain performance data.

```
# time PATH=/usr/lib64/genwqe:$PATH \
tar cfz linux-src.hw.tar.gz linux-src
real 0m1.323s
user 0m0.242s
sys 0m1.023s
```
Results

In the example, the accelerated operation is significantly faster. The hardware-compressed data is slightly larger than the software-compressed version of the same data.

GenWQE hardware-acceleration for IBM Java

IBM Java version 7.1 or later can use the GenWQE hardware-accelerated zlib.

To activate the GenWQE hardware-accelerated zlib for IBM Java, you must set environment parameters. See the documentation for your Java version to find out which settings are required.

Note: Any values that you set for the environment variables override the default settings for the GenWQE user space components (see “Environment variables” on page 362).

Exploring the GenWQE setup

You might want to ensure that your GenWQE setup works as intended.

- “Listing your GenWQE accelerator cards”
- “Checking the GenWQE device driver setup”
- “Confirming that the accelerator hardware can be reached” on page 366

Listing your GenWQE accelerator cards

Use the `lspci` command to list the available GenWQE accelerator cards.

Procedure

1. Issue the `lspci` command and look for GenWQE.

   Example:

   ```
   # lspci |grep GenWQE
   0002:00:00.0 Processing accelerators: IBM GenWQE Accelerator Adapter
   ```

2. Issue the `lspci` command with the verbose option to display details about a particular card.

   Example:

   ```
   # lspci -vs 0002:00:00.0
   0002:00:00.0 Processing accelerators: IBM GenWQE Accelerator Adapter
   Subsystem: IBM GenWQE Accelerator Adapter
   Physical Slot: 0000000f
   Flags: bus master, fast devsel, latency 0, IRQ 3
   Memory at 8002000000000000 (64-bit, prefetchable) [disabled] [size=128M]
  Capabilities: [50] MSI: Enable+ Count=1 Maskable- 64bit+
   Capabilities: [80] Express Endpoint, MSI 00
   Capabilities: [100] Alternative Routing-ID Interpretation (ARI)
   Kernel driver in use: genwqe
   Kernel modules: genwqe_card
   ```

Checking the GenWQE device driver setup

Perform these tasks if GenWQE does not work as expected.
Procedure

1. Confirm that the device driver is loaded.

```
# lsmod | grep genwqe
genwqe_card 88997 0
crc_itu_t 1910 1 genwqe_card
```

If the `genwqe_card` module is not listed in the command output, load it with `modprobe`.

```
# modprobe genwqe_card
```

The `genwqe_card` module does not have module parameters.

2. Confirm that GenWQE device nodes exist and that the nodes have the required permissions.
   The nodes must grant read and write permissions to all users, for example:

```
# ls -l /dev/genwqe*
crwrwrw 1 root root 249, 0 Jun 30 10:01 /dev/genwqe0_card
crwrwrw 1 root root 248, 0 Jun 30 10:01 /dev/genwqe1_card
```

If the permissions are not `crwrwrw`, create a file `/etc/udev/rules.d/52-genwqedevice.rules` with this rule as its content:

```
KERNEL="genwqe*", MODE="0666"
```

The new rule takes effect next time the GenWQE device driver is loaded.

**Tip:** Use the `chmod` command to temporarily set the permissions.

What to do next

You can find debug information in the Linux source tree at `/sys/kernel/debug/genwqe` and at `/sys/class/genwqe`.

Confirming that the accelerator hardware can be reached

The `genwqe_echo` command is similar to a `ping` command.

Before you begin

The `genwqe_echo` command is included in the `genwqe-tools` RPM (see "Installing the GenWQE hardware-accelerated zlib" on page 362).

Procedure

Issue a command of this form to confirm that you can reach the accelerator hardware.

```
# genwqe_echo -AGENWQE -C <n> -c <m>
```

In the command, `<n>` is the index number of the card and `<m>` is a positive integer that specifies how many requests are sent to the card. The `-AGENWQE` parameter ensures that the correct, PCIe-attached, accelerator card is used.
**Example:** The following command sends four requests to the card with device node `/dev/genwqe1_card`:

```
# genwqe_echo -AGENWQE -C 1 -c 4
1 x 33 bytes from UNIT #1: echo_req time=37.0 usec
1 x 33 bytes from UNIT #1: echo_req time=19.0 usec
1 x 33 bytes from UNIT #1: echo_req time=23.0 usec
1 x 33 bytes from UNIT #1: echo_req time=18.0 usec
--- UNIT #1 echo statistics ---
4 packets transmitted, 4 received, 0 lost, 0% packet loss
```

See the `genwqe_echo` man page for other command options.

**External programming interfaces**

The GenWQE hardware-accelerated zlib implements a large subset of the original software zlib.

For information about programming against the GenWQE hardware-accelerated zlib, see the section about implemented zlib functions in *Accelerated Data Compressing using the GenWQE Linux Driver and Corsa FPGA PCIe card*.

To obtain this document, go to the developerWorks website at [www.ibm.com/developerworks/community/files/app](http://www.ibm.com/developerworks/community/files/app) and search for “genwqe”.

Part 6. z/VM virtual server integration

<table>
<thead>
<tr>
<th>Chapter 29. z/VM concepts</th>
<th>371</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance monitoring for z/VM guest virtual machines</td>
<td>371</td>
</tr>
<tr>
<td>Cooperative memory management background</td>
<td>373</td>
</tr>
<tr>
<td>Linux guest relocation</td>
<td>374</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 30. Writing kernel APPLDATA records</th>
<th>375</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting up the APPLDATA record support</td>
<td>375</td>
</tr>
<tr>
<td>Generating APPLDATA monitor records</td>
<td>375</td>
</tr>
<tr>
<td>APPLDATA monitor record layout</td>
<td>377</td>
</tr>
<tr>
<td>Programming interfaces</td>
<td>380</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 31. Writing z/VM monitor records</th>
<th>381</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting up the z/VM *MONITOR record writer device driver</td>
<td>381</td>
</tr>
<tr>
<td>Working with the z/VM *MONITOR record writer</td>
<td>382</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 32. Reading z/VM monitor records</th>
<th>385</th>
</tr>
</thead>
<tbody>
<tr>
<td>What you should know about the z/VM *MONITOR record reader device driver</td>
<td>385</td>
</tr>
<tr>
<td>Setting up the z/VM *MONITOR record reader device driver</td>
<td>386</td>
</tr>
<tr>
<td>Working with the z/VM *MONITOR record reader</td>
<td>388</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 33. z/VM recording device driver</th>
<th>391</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>391</td>
</tr>
<tr>
<td>What you should know about the z/VM recording device driver</td>
<td>391</td>
</tr>
<tr>
<td>Setting up the z/VM recording device driver</td>
<td>392</td>
</tr>
</tbody>
</table>

These device drivers and features help you to effectively run and manage a z/VM-based virtual Linux server farm.

Newest version

You can find the newest version of this book at
www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_distlibs.html

Restrictions

For prerequisites and restrictions see the IBM Z architecture specific information in the Red Hat Enterprise Linux 7.5 release notes at
https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux
Chapter 29. z/VM concepts

The z/VM performance monitoring and cooperative memory management concepts help you to understand how the different components interact with Linux.

Performance monitoring for z/VM guest virtual machines

You can monitor the performance of z/VM guest virtual machines and their guest operating systems with performance monitoring tools on z/VM or on Linux.

These tools can be your own, IBM tools such as the Performance Toolkit for VM, or third-party tools. The guests being monitored require agents that write monitor data.

Monitoring on z/VM

z/VM monitoring tools must read performance data. For monitoring Linux instances, this data is APPLDATA monitor records.

Linux instances must write these records for the tool to read, as shown in Figure 71.

![Diagram showing data flow](image)

**Figure 71. Linux instances write APPLDATA records for performance monitoring tools**

Both user space applications and the Linux kernel can write performance data to APPLDATA records. Applications use the monwriter device driver to write APPLDATA records. The Linux kernel can be configured to collect system level data such as memory, CPU usage, and network-related data, and write it to data records.

For file system size data, there is a command, `mon_fsstatd`. This user space tool uses the monwriter device driver to write file system size information as defined records.

For process data, there is a command, `mon_procd`. This user space tool uses the monwriter device driver to write system information as defined records.
In summary, Red Hat Enterprise Linux 7.5 for z Systems supports writing and collecting performance data as follows:

- The Linux kernel can write z/VM monitor data for Linux instances, see Chapter 30, “Writing kernel APPLDATA records,” on page 375.
- Linux applications that run on z/VM guests can write z/VM monitor data, see Chapter 31, “Writing z/VM monitor records,” on page 381.
- You can collect monitor file system size information, see “mon_fsstatd – Monitor z/VM guest file system size” on page 597.
- You can collect system information about up to 100 concurrently running processes, see “mon_procd – Monitor Linux on z/VM” on page 601.

**Monitoring on Linux**

A Linux instance can read the monitor data by using the monreader device driver.

**Figure 72** illustrates a Linux instance that is set up to read the monitor data. You can use an existing monitoring tool or write your own software.

In summary, Red Hat Enterprise Linux 7.5 for z Systems supports reading performance data in the form of read access to z/VM monitor data for Linux instances. See Chapter 32, “Reading z/VM monitor records,” on page 385 for more details.

**Further information**

Several z/VM publications include information about monitoring.

- See z/VM Getting Started with Linux on System z, SC24-6194, the chapter about monitoring performance for information about using the CP Monitor and the Performance Toolkit for VM.
- See z/VM Saved Segments Planning and Administration, SC24-6229 for general information about DCSSs (z/VM keeps monitor records in a DCSS).
- See z/VM Performance, SC24-6208 for information about creating a monitor DCSS.
- See z/VM CP Commands and Utilities Reference, SC24-6175 for information about the CP commands that are used in the context of DCSSs and for controlling the z/VM monitor system service.
Cooperative memory management background

Cooperative memory management (CMM, or "cmm1") dynamically adjusts the memory available to Linux.

For information about setting up CMM, see Chapter 38, “Cooperative memory management,” on page 421.

In a virtualized environment it is common practice to give the virtual machines more memory than is actually available to the hypervisor. Linux tends to use all of its available memory. As a result, the hypervisor (z/VM) might start swapping.

To avoid excessive z/VM swapping, the memory available to Linux can be reduced. CMM allocates pages to page pools that make the pages unusable to Linux. There are two such page pools as shown in Figure 73.

z/VM memory

There are two page pools:

A static page pool
The page pool is controlled by a resource manager that changes the pool size at intervals according to guest activity and overall memory usage on z/VM (see Figure 74 on page 374).
A timed page pool

Pages are released from this pool at a speed that is set in the release rate (see Figure 75). According to guest activity and overall memory usage on z/VM, a resource manager adds pages at intervals. If no pages are added and the release rate is not zero, the pool empties.

The external resource manager that controls the pools can be the z/VM resource monitor (VMRM) or a third-party systems management tool.

VMRM controls the pools over a message interface. Setting up the external resource manager is beyond the scope of this information. For more details, see the chapter on VMRM in z/VM Performance, SC24-6208.

Third-party tools can provide a Linux daemon that receives commands for the memory allocation through TCP/IP. The daemon, in turn, uses the procfs-based interface. You can use the procfs interface to read the pool sizes. These values are useful diagnostic data.

Linux guest relocation

Information about guest relocations is stored in the s390 debug feature (s390dbf).

You can access this information in a kernel dump or from a running Linux instance. For more information, see Linux on z Systems Troubleshooting, SC34-2612.
Chapter 30. Writing kernel APPLDATA records

z/VM is a convenient point for collecting z/VM guest performance data and statistics for an entire server farm. Linux instances can export such data to z/VM by using APPLDATA monitor records.

z/VM regularly collects these records. The records are then available to z/VM performance monitoring tools.

A virtual CPU timer on the Linux instance to be monitored controls when data is collected. The timer accounts for only busy time to avoid unnecessarily waking up an idle guest. The APPLDATA record support comprises several modules. A base module provides an intra-kernel interface and the timer function. The intra-kernel interface is used by data gathering modules that collect actual data and determine the layout of a corresponding APPLDATA monitor record (see “APPLDATA monitor record layout” on page 377).

For an overview of performance monitoring support, see “Performance monitoring for z/VM guest virtual machines” on page 371.

Setting up the APPLDATA record support

You must enable your z/VM guest virtual machine for data gathering and load the APPLDATA record support modules.

Procedure
1. On z/VM, ensure that the user directory of the guest virtual machine includes the option APPLMON.
2. On Linux, use the `modprobe` command to load any required modules.

```
modprobe appldata_mem appldata_os appldata_net_sum
```

where appldata_mem, appldata_os, and appldata_net_sum are the modules for gathering memory-related data, operating system-related data, and network-related data.

See the `modprobe` man page for command details.

Generating APPLDATA monitor records

You can set the timer interval and enable or disable data collection.

You control the monitor stream support through the procfs. APPLDATA monitor records are produced if both a particular data gathering module and the monitoring support in general are enabled.
Enabling or disabling the support

Use the procfs timer attribute to enable or disable the monitoring support.

**Procedure**

To read the current setting, issue:

```
# cat /proc/sys/appldata/timer
```

To enable the monitoring support issue:

```
# echo 1 > /proc/sys/appldata/timer
```

To disable the monitoring support issue:

```
# echo 0 > /proc/sys/appldata/timer
```

Activating or deactivating individual data-gathering modules

Each data-gathering module has a procfs entry that contains a value 1 if the module is active and 0 if the module is inactive.

**About this task**

The following procfs entries control the data-gathering modules:

- `/proc/sys/appldata/mem` for the memory data-gathering module
- `/proc/sys/appldata/os` for the CPU data-gathering module
- `/proc/sys/appldata/net_sum` for the net data-gathering module

To check whether a module is active look at the content of the corresponding procfs entry.

**Procedure**

To activate a data-gathering module write 1 to the corresponding procfs entry. To deactivate a data-gathering module write 0 to the corresponding procfs entry.

Issue a command of this form:

```
# echo <flag> > /proc/sys/appldata/<data_type>
```

where `<data_type>` is one of mem, os, or net_sum.

**Note:** An active data-gathering module produces APPLDATA monitor records only if the monitoring support is enabled (see “Enabling or disabling the support”).

**Example**

To find out whether memory data-gathering is active, issue:

```
# cat /proc/sys/appldata/mem
0
```
In the example, memory data-gathering is off. To activate memory data-gathering, issue:

```
# echo 1 > /proc/sys/appldata/mem
```

To deactivate the memory data-gathering module, issue:

```
# echo 0 > /proc/sys/appldata/mem
```

### Setting the sampling interval

You can set the time that lapses between consecutive data samples.

#### About this task

The time that you set is measured by the virtual CPU timer. Because the virtual timer slows down as the guest idles, the sampling interval in real time can be considerably longer than the value you set.

The value in `/proc/sys/appldata/interval` is the sample interval in milliseconds. The default sample interval is 10 000 ms.

#### Procedure

To read the current value, issue:

```
# cat /proc/sys/appldata/interval
```

To set the sample interval to a different value, write the new value (in milliseconds) to `/proc/sys/appldata/interval`. Issue a command of this form:

```
# echo <interval> > /proc/sys/appldata/interval
```

where `<interval>` is the new sample interval in milliseconds. The specification must be in the range 1 - 2147483647, where 2,147,483,647 = $2^{31} - 1$.

#### Example

To set the sampling interval to 20 s (20000 ms), issue:

```
# echo 20000 > /proc/sys/appldata/interval
```

### APPLDATA monitor record layout

Each of the data gathering modules writes a different type of record.

- Memory data (see Table 47 on page 378)
- Processor data (see Table 48 on page 378)
- Networking (see Table 49 on page 379)

z/VM can identify the records by their unique product ID. The product ID is an EBCDIC string of this form: “LINUXKRNL<record ID>260100”. The `<record ID>` is treated as a byte value, not a string.
The records contain data of the following types:

**u32**  unsigned 4 byte integer

**u64**  unsigned 8 byte integer

*Table 47. APPLDATA_MEM_DATA record (Record ID 0x01)*

<table>
<thead>
<tr>
<th>Offset (Decimal)</th>
<th>Offset (Hex)</th>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0x0</td>
<td>u64</td>
<td>timestamp</td>
<td>TOD time stamp that is generated on the Linux side after record update</td>
</tr>
<tr>
<td>8</td>
<td>0x8</td>
<td>u32</td>
<td>sync_count_1</td>
<td>After z/VM collected the record data, sync_count_1 and sync_count_2 must be the same. Otherwise, the record was updated on the Linux side while z/VM was collecting the data. As a result, the data might be inconsistent.</td>
</tr>
<tr>
<td>12</td>
<td>0xC</td>
<td>u32</td>
<td>sync_count_2</td>
<td>See sync_count_1.</td>
</tr>
<tr>
<td>16</td>
<td>0x10</td>
<td>u64</td>
<td>pgpgin</td>
<td>Data that was read from disk (in KB)</td>
</tr>
<tr>
<td>24</td>
<td>0x18</td>
<td>u64</td>
<td>pgpgcout</td>
<td>Data that was written to disk (in KB)</td>
</tr>
<tr>
<td>32</td>
<td>0x20</td>
<td>u64</td>
<td>pswpin</td>
<td>Pages that were swapped in</td>
</tr>
<tr>
<td>40</td>
<td>0x28</td>
<td>u64</td>
<td>pswpout</td>
<td>Pages that were swapped out</td>
</tr>
<tr>
<td>48</td>
<td>0x30</td>
<td>u64</td>
<td>sharedram</td>
<td>Shared RAM in KB, set to 0</td>
</tr>
<tr>
<td>56</td>
<td>0x38</td>
<td>u64</td>
<td>totalram</td>
<td>Total usable main memory size in KB</td>
</tr>
<tr>
<td>64</td>
<td>0x40</td>
<td>u64</td>
<td>freeram</td>
<td>Available memory size in KB</td>
</tr>
<tr>
<td>72</td>
<td>0x48</td>
<td>u64</td>
<td>totalhigh</td>
<td>Total high memory size in KB</td>
</tr>
<tr>
<td>80</td>
<td>0x50</td>
<td>u64</td>
<td>freehigh</td>
<td>Available high memory size in KB</td>
</tr>
<tr>
<td>88</td>
<td>0x58</td>
<td>u64</td>
<td>bufferram</td>
<td>Memory that was reserved for raw disk blocks, corresponding to &quot;Buffers&quot; from /proc/meminfo, in KB</td>
</tr>
<tr>
<td>96</td>
<td>0x60</td>
<td>u64</td>
<td>cached</td>
<td>Size of used cache, including &quot;Cached&quot; and &quot;SwapCached&quot; from /proc/meminfo, in KB</td>
</tr>
<tr>
<td>104</td>
<td>0x68</td>
<td>u64</td>
<td>totalswap</td>
<td>Total swap space size in KB</td>
</tr>
<tr>
<td>112</td>
<td>0x70</td>
<td>u64</td>
<td>freeswap</td>
<td>Free swap space in KB</td>
</tr>
<tr>
<td>120</td>
<td>0x78</td>
<td>u64</td>
<td>pgalloc</td>
<td>Page allocations</td>
</tr>
<tr>
<td>128</td>
<td>0x80</td>
<td>u64</td>
<td>pgfault</td>
<td>Page faults (major+minor)</td>
</tr>
<tr>
<td>136</td>
<td>0x88</td>
<td>u64</td>
<td>pgmajfault</td>
<td>Page faults (major only)</td>
</tr>
</tbody>
</table>

*Table 48. APPLDATA_OS_DATA record (Record ID 0x02)*

<table>
<thead>
<tr>
<th>Offset (Decimal)</th>
<th>Offset (Hex)</th>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0x0</td>
<td>u64</td>
<td>timestamp</td>
<td>TOD time stamp that is generated on the Linux side after record update</td>
</tr>
<tr>
<td>8</td>
<td>0x8</td>
<td>u32</td>
<td>sync_count_1</td>
<td>After z/VM collected the record data, sync_count_1 and sync_count_2 must be the same. Otherwise, the record was updated on the Linux side while z/VM was collecting the data. As a result, the data might be inconsistent.</td>
</tr>
<tr>
<td>12</td>
<td>0xC</td>
<td>u32</td>
<td>sync_count_2</td>
<td>See sync_count_1.</td>
</tr>
<tr>
<td>16</td>
<td>0x10</td>
<td>u32</td>
<td>nr_cpus</td>
<td>Number of virtual CPUs.</td>
</tr>
</tbody>
</table>
### Table 48. APPLDATA_OS_DATA record (Record ID 0x02)  (continued)

<table>
<thead>
<tr>
<th>Offset (Decimal)</th>
<th>Offset (Hex)</th>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0x14</td>
<td>u32</td>
<td>per_cpu_size</td>
<td>Size of the per_cpu_data for each CPU (= 36).</td>
</tr>
<tr>
<td>24</td>
<td>0x18</td>
<td>u32</td>
<td>cpu_offset</td>
<td>Offset of the first per_cpu_data (= 52).</td>
</tr>
<tr>
<td>28</td>
<td>0x1C</td>
<td>u32</td>
<td>nr_running</td>
<td>Number of runnable threads.</td>
</tr>
<tr>
<td>32</td>
<td>0x20</td>
<td>u32</td>
<td>nr_threads</td>
<td>Number of threads.</td>
</tr>
<tr>
<td>36</td>
<td>0x24</td>
<td>3 × u32</td>
<td>avenrun[3]</td>
<td>Average number of running processes during the last 1 (1st value), 5 (2nd value) and 15 (3rd value) minutes. These values are 'fake fix-point', each value is composed of a 10-bit integer and an 11-bit fractional part. See note[1] at the end of this table.</td>
</tr>
<tr>
<td>48</td>
<td>0x30</td>
<td>u32</td>
<td>nr_iowait</td>
<td>Number of blocked threads (waiting for I/O).</td>
</tr>
<tr>
<td>52</td>
<td>0x34</td>
<td>See note[2]</td>
<td>per_cpu_data</td>
<td>Time that is spent in user, kernel, idle, nice, etc for every CPU. See note[3] at the end of this table.</td>
</tr>
<tr>
<td>52</td>
<td>0x34</td>
<td>u32</td>
<td>per_cpu_user</td>
<td>Timer ticks that were spent in user mode.</td>
</tr>
<tr>
<td>56</td>
<td>0x38</td>
<td>u32</td>
<td>per_cpu_nice</td>
<td>Timer ticks that were spent with modified priority.</td>
</tr>
<tr>
<td>60</td>
<td>0x3C</td>
<td>u32</td>
<td>per_cpu_system</td>
<td>Timer ticks that were spent in kernel mode.</td>
</tr>
<tr>
<td>64</td>
<td>0x40</td>
<td>u32</td>
<td>per_cpu_idle</td>
<td>Timer ticks that were spent in idle mode.</td>
</tr>
<tr>
<td>68</td>
<td>0x44</td>
<td>u32</td>
<td>per_cpu_irq</td>
<td>Timer ticks that were spent in interrupts.</td>
</tr>
<tr>
<td>72</td>
<td>0x48</td>
<td>u32</td>
<td>per_cpu_softirq</td>
<td>Timer ticks that were spent in softirqs.</td>
</tr>
<tr>
<td>76</td>
<td>0x4C</td>
<td>u32</td>
<td>per_cpu_iowait</td>
<td>Timer ticks that were spent while waiting for I/O.</td>
</tr>
<tr>
<td>80</td>
<td>0x50</td>
<td>u32</td>
<td>per_cpu_steal</td>
<td>Timer ticks &quot;stolen&quot; by hypervisor.</td>
</tr>
<tr>
<td>84</td>
<td>0x54</td>
<td>u32</td>
<td>cpu_id</td>
<td>The number of this CPU.</td>
</tr>
</tbody>
</table>

**Note:**
1. The following C-Macros are used inside Linux to transform these into values with two decimal places:
   ```c
   #define LOAD_INT(x) (((x) >> 11)
   #define LOAD_FRAC(x) LOAD_INT(((x) & (((1 << 11) - 1)) * 100)
   ```
2. `nr_cpus * per_cpu_size`
3. `per_cpu_user` through `cpu_id` are repeated for each CPU

### Table 49. APPLDATA_NET_SUM_DATA record (Record ID 0x03)

<table>
<thead>
<tr>
<th>Offset (Decimal)</th>
<th>Offset (Hex)</th>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0x0</td>
<td>u64</td>
<td>timestamp</td>
<td>TOD time stamp that is generated on the Linux side after record update</td>
</tr>
</tbody>
</table>
Table 49. APPLDATA_NET_SUM_DATA record (Record ID 0x03) (continued)

<table>
<thead>
<tr>
<th>Offset (Decimal)</th>
<th>Offset (Hex)</th>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0x8</td>
<td>u32</td>
<td>sync_count_1</td>
<td>After z/VM collected the record data, sync_count_1 and sync_count_2 must be the same. Otherwise, the record was updated on the Linux side while z/VM was collecting the data. As a result, the data might be inconsistent.</td>
</tr>
<tr>
<td>12</td>
<td>0xC</td>
<td>u32</td>
<td>sync_count_2</td>
<td>See sync_count_1.</td>
</tr>
<tr>
<td>16</td>
<td>0x10</td>
<td>u32</td>
<td>nr_interfaces</td>
<td>Number of interfaces being monitored</td>
</tr>
<tr>
<td>20</td>
<td>0x14</td>
<td>u32</td>
<td>padding</td>
<td>Unused. The next value is 64-bit aligned, so these 4 byte would be padded out by compiler</td>
</tr>
<tr>
<td>24</td>
<td>0x18</td>
<td>u64</td>
<td>rx_packets</td>
<td>Total packets that were received</td>
</tr>
<tr>
<td>32</td>
<td>0x20</td>
<td>u64</td>
<td>tx_packets</td>
<td>Total packets that were transmitted</td>
</tr>
<tr>
<td>40</td>
<td>0x28</td>
<td>u64</td>
<td>rx_bytes</td>
<td>Total bytes that were received</td>
</tr>
<tr>
<td>48</td>
<td>0x30</td>
<td>u64</td>
<td>tx_bytes</td>
<td>Total bytes that were transmitted</td>
</tr>
<tr>
<td>56</td>
<td>0x38</td>
<td>u64</td>
<td>rx_errors</td>
<td>Number of bad packets that were received</td>
</tr>
<tr>
<td>64</td>
<td>0x40</td>
<td>u64</td>
<td>tx_errors</td>
<td>Number of packet transmit problems</td>
</tr>
<tr>
<td>72</td>
<td>0x48</td>
<td>u64</td>
<td>rx_dropped</td>
<td>Number of incoming packets that were dropped because of insufficient space in Linux buffers</td>
</tr>
<tr>
<td>80</td>
<td>0x50</td>
<td>u64</td>
<td>tx_dropped</td>
<td>Number of outgoing packets that were dropped because of insufficient space in Linux buffers</td>
</tr>
<tr>
<td>88</td>
<td>0x58</td>
<td>u64</td>
<td>collisions</td>
<td>Number of collisions while transmitting</td>
</tr>
</tbody>
</table>

### Programming interfaces

The monitor stream support base module exports two functions.

**Application programmers:** This information is intended for those who want to program against the monitor stream.

- appldata_register_ops() to register data-gathering modules
- appldata_unregister_ops() to undo the registration of data-gathering modules

Both functions receive a pointer to a struct appldata_ops as parameter. Additional data-gathering modules that want to plug into the base module must provide this data structure. You can find the definition of the structure and the functions in arch/s390/appldata/appldata.h in the Linux source tree.

See "APPLDATA monitor record layout" on page 377 for an example of APPLDATA data records that are to be sent to z/VM.

**Tip:** Include the timestamp, sync_count_1, and sync_count_2 fields at the beginning of the record as shown for the existing APPLDATA record formats.
Chapter 31. Writing z/VM monitor records

Applications can use the monitor stream application device driver to write z/VM monitor APPLDATA records to the z/VM *MONITOR stream.

For an overview of performance monitoring support, see “Performance monitoring for z/VM guest virtual machines” on page 371.

The monitor stream application device driver interacts with the z/VM monitor APPLDATA facilities for performance monitoring. A better understanding of these z/VM facilities might help when you are using this device driver. See z/VM Performance, SC24-6208 for information about monitor APPLDATA.

The monitor stream application device driver provides the following functions:

- An interface to the z/VM monitor stream.
- A means of writing z/VM monitor APPLDATA records.

Setting up the z/VM *MONITOR record writer device driver

You must load the monwriter module on Linux and set up your guest virtual machine for monitor records on z/VM.

Module parameters

You can configure the monitor stream application device driver when you are loading the device driver module, monwriter.

```
Monitor stream application device driver module parameter syntax

modprobe monwriter max_bufs=255
```

where `<numbufs>` is the maximum number of monitor sample and configuration data buffers that can exist in the Linux guest at one time. The default is 255.

Example

To load the monwriter module and set the maximum number of buffers to 400, use the following command:

```
# modprobe monwriter max_bufs=400
```

Setting up the z/VM guest virtual machine

You must enable your z/VM guest virtual machine to write monitor records and configure the z/VM system to collect these records.
Procedure

Perform these steps:

1. Set this option in the z/VM user directory entry of the virtual machine in which the application that uses this device driver is to run:
   - OPTION APPLMON

2. Issue the following CP commands to have CP collect the respective types of monitor data:
   - MONITOR SAMPLE ENABLE APPLDATA ALL
   - MONITOR EVENT ENABLE APPLDATA ALL

   You can log in to the z/VM console to issue the CP commands. These commands must be preceded with #CP. Alternatively, you can use the `vmcp` command for issuing CP commands from your Linux instance.

   See z/VM CP Commands and Utilities Reference, SC24-6175 for information about the CP MONITOR command.

Working with the z/VM *MONITOR record writer

The monitor stream application device driver uses the z/VM CP instruction DIAG X'DC' to write to the z/VM monitor stream. Monitor data must be preceded by a data structure, monwrite_hdr.

See z/VM CP Programming Services, SC24-6179 for more information about the DIAG X'DC' instruction and the different monitor record types (sample, config, event).

The application writes monitor data by passing a monwrite_hdr structure that is followed by monitor data. The only exception is the STOP function, which requires no monitor data. The monwrite_hdr structure, as described in monwriter.h, is filled in by the application. The structure includes the DIAG X'DC' function to be performed, the product identifier, the header length, and the data length.

All records that are written to the z/VM monitor stream begin with a product identifier. This device driver uses the product ID. The product ID is a 16-byte structure of the form pppppppppfnvrrmm, where:

pppppppp

is a fixed ASCII string, for example, LNXAPPL.

ff

is the application number (hexadecimal number). This number can be chosen by the application. You can reduce the chance of conflicts with other applications, by requesting an application number from the IBM z/VM Performance team at


n

is the record number as specified by the application

vv, rr, and mm

can also be specified by the application. A possible use is to specify version, release, and modification level information, allowing changes to a certain record number when the layout is changed, without changing the record number itself.
The first 7 bytes of the structure (LNXAPPL) are filled in by the device driver when it writes the monitor data record to the CP buffer. The last 9 bytes contain information that is supplied by the application on the write() call when writing the data.

The monwrite_hdr structure that must be written before any monitor record data is defined as follows:

```c
/* the header the app uses in its write() data */
struct monwrite_hdr {
    unsigned char mon_function;
    unsigned short applid;
    unsigned char record_num;
    unsigned short version;
    unsigned short release;
    unsigned short mod_level;
    unsigned short datalen;
    unsigned char hdrlen;
} __attribute__((packed));
```

The following function code values are defined:

```c
/* mon_function values */
#define MONWRITE_START_INTERVAL 0x00 /* start interval recording */
#define MONWRITE_STOP_INTERVAL 0x01 /* stop interval or config recording */
#define MONWRITE_GEN_EVENT 0x02 /* generate event record */
#define MONWRITE_START_CONFIG 0x03 /* start configuration recording */
```

**Writing data and stopping data writing**

Applications use the open(), write(), and close() calls to work with the z/VM monitor stream.

Before an application can write monitor records, it must issue open() to open the device driver. Then, the application must issue write() calls to start or stop the collection of monitor data and to write any monitor records to buffers that CP can access.

When the application has finished writing monitor data, it must issue close() to close the device driver.

**Using the monwrite_hdr structure**

The structure monwrite_hdr is used to pass DIAG x’DC’ functions and the application-defined product information to the device driver on write() calls.

When the application calls write(), the data it is writing consists of one or more monwrite_hdr structures. Each structure is followed by monitor data. The only exception is the STOP function, which is not followed by data.

The application can write to one or more monitor buffers. A new buffer is created by the device driver for each record with a unique product identifier. To write new data to an existing buffer, an identical monwrite_hdr structure must precede the new data on the write() call.

The monwrite_hdr structure also includes a field for the header length, which is useful for calculating the data offset from the beginning of the header. There is also a field for the data length, which is the length of any monitor data that follows. See /usr/include/asm/monwriter.h for the definition of the monwrite_hdr structure.
Chapter 32. Reading z/VM monitor records

Monitoring software on Linux can access z/VM guest data through the z/VM *MONITOR record reader device driver.

z/VM uses the z/VM monitor system service (*MONITOR) to collect monitor records from agents on its guests. z/VM writes the records to a discontiguous saved segment (DCSS). See z/VM Saved Segments Planning and Administration, SC24-6229 for general information about DCSSs.

The z/VM *MONITOR record reader device driver uses IUCV to connect to *MONITOR and accesses the DCSS as a character device.

For an overview of performance monitoring support, see “Performance monitoring for z/VM guest virtual machines” on page 371.

The z/VM *MONITOR record reader device driver supports the following devices and functions:

- Read access to the z/VM *MONITOR DCSS.
- Reading *MONITOR records for z/VM.
- Access to the kernel APPLDATA records from the Linux monitor stream (see Chapter 30, “Writing kernel APPLDATA records,” on page 375).

What you should know about the z/VM *MONITOR record reader device driver

The data that is collected by *MONITOR depends on the setup of the monitor stream service.

The z/VM *MONITOR record reader device driver only reads data from the monitor DCSS; it does not control the system service.

z/VM supports only one monitor DCSS. All monitoring software that requires monitor records from z/VM uses the same DCSS to read *MONITOR data. Usually, a DCSS called MONDCSS is already defined and used by existing monitoring software.

If a monitor DCSS is already defined, you must use it. To find out whether a monitor DCSS exists, issue the following CP command from a z/VM guest virtual machine with privilege class E:

```
q monitor
```

The command output also shows the name of the DCSS.
Device node

Red Hat Enterprise Linux 7.5 creates a device node for you using udev. The device node is called /dev/monreader and is a miscellaneous character device that you can use to access the monitor DCSS.

Further information

- See z/VM Saved Segments Planning and Administration, SC24-6229 for general information about DCSSs.
- See z/VM Performance, SC24-6208 for information about creating a monitor DCSS.
- See z/VM CP Commands and Utilities Reference, SC24-6175 for information about the CP commands that are used in the context of DCSSs and for controlling the z/VM monitor system service.
- For the layout of the monitor records, go to www.ibm.com/vm/pubs/ctlblk.html and click the link to the monitor record format for your z/VM version. Also, see Chapter 30, “Writing kernel APPLDATA records,” on page 375.

Setting up the z/VM *MONITOR record reader device driver

You must set up Linux and the z/VM guest virtual machine for accessing an existing monitor DCSS with the z/VM *MONITOR record reader device driver.

Before you begin

Some of the CP commands you use for setting up the z/VM *MONITOR record reader device driver require class E authorization.

Setting up the monitor system service and the monitor DCSS on z/VM is beyond the scope of this information. See z/VM Performance, SC24-6208 for information about creating a monitor DCSS.

Providing the required user directory entries for your z/VM guest

The z/VM guest where your Linux instance is to run must be permitted to establish an IUCV connection to the z/VM *MONITOR system service.

About this task

See z/VM CP Commands and Utilities Reference, SC24-6175 for information about the CP commands that are used in the context of DCSSs and for controlling the z/VM monitor system service.

Procedure

Ensure that the guest entry in the user directory includes the following statement:

IUCV *MONITOR

If the DCSS is restricted, you also need this statement:

NAMESAVE <dcss>

where <dcss> is the name of the DCSS that is used for the monitor records. You can find out the name of an existing monitor DCSS by issuing the following CP command from a z/VM guest virtual machine with privilege class E:
Assuring that the DCSS is addressable for your Linux instance

The DCSS address range must not overlap with the storage of your z/VM guest virtual machine.

Procedure

To find out the start and end address of the DCSS, issue the following CP command from a z/VM guest virtual machine with privilege class E:

q nss map

The output gives you the start and end addresses of all defined DCSSs in units of 4-kilobyte pages. For example:

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>00:</td>
<td>FILE</td>
<td>FILENAME</td>
<td>FILETYPE</td>
<td>MINSIZE</td>
<td>BEGPAG</td>
<td>ENDPAG</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>00:</td>
<td>0011</td>
<td>MONDCSS</td>
<td>CPDCSS</td>
<td>N/A</td>
<td>09000</td>
<td>097FF</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What to do next

If the DCSS overlaps with the guest storage, follow the procedure in "Avoiding overlaps with your guest storage" on page 403.

Specifying the monitor DCSS name

Specify the DCSS name as a module parameter when you load the module.

About this task

By default, the z/VM *MONITOR record reader device driver assumes that the monitor DCSS on z/VM is called MONDCSS. If you want to use a different DCSS name, you must specify it.

Load the monitor read support module with modprobe to assure that any other required modules are also loaded. You need IUCV support if you want to use the monitor read support.

```
modprobe monreader
```

where `<dcss>` is the name of the DCSS that z/VM uses for the monitor records.
Example

To load the monitor read support module and specify MYDCSS as the DCSS issue:

```
modprobe monreader mondcss=mydcss
```

---

**Working with the z/VM *MONITOR record reader**

You can open the z/VM *MONITOR record character device to read records from it.

This section describes how to work with the monitor read support.

- “Opening and closing the character device”
- “Reading monitor records”

**Opening and closing the character device**

Only one user can open the character device at any one time. Once you have opened the device, you must close it to make it accessible to other users.

**About this task**

The open function can fail (return a negative value) with one of the following values for errno:

- **EBUSY**
  The device has already been opened by another user.

- **EIO**
  No IUCV connection to the z/VM MONITOR system service could be established. An error message with an IPUSER SEVER code is printed into syslog. See z/VM Performance, SC24-6208 for details about the codes.

Once the device is opened, incoming messages are accepted and account for the message limit. If you keep the device open indefinitely, expect to eventually reach the message limit (with error code EOVERFLOW).

**Reading monitor records**

You can either read in non-blocking mode with polling, or you can read in blocking mode without polling.

**About this task**

Reading from the device provides a 12-byte monitor control element (MCE), followed by a set of one or more contiguous monitor records (similar to the output of the CMS utility MONWRITE without the 4 K control blocks). The MCE contains information about:

- The type of the following record set (sample/event data)
- The monitor domains contained within it
- The start and end address of the record set in the monitor DCSS

The start and end address can be used to determine the size of the record set. The end address is the address of the last byte of data. The start address is needed to handle "end-of-frame" records correctly (domain 1, record 13), that is, it can be used to determine the record start offset relative to a 4 K page (frame) boundary.
See "Appendix A: *MONITOR" in z/VM Performance, SC24-6208 for a description of the monitor control element layout. For the layout of the monitor records go to www.ibm.com/ibm/ and click the link to the monitor record format for your z/VM version. Also see Chapter 30, “Writing kernel APPLDATA records,” on page 375.

The layout of the data stream that is provided by the monreader device is as follows:

...<0 byte read> <first MCE> \ <first set of records> | ... ... - data set <last MCE> |<last set of records> / <0 byte read> ...

There might be more than one combination of MCE and a corresponding record set within one data set. The end of each data set is indicated by a successful read with a return value of 0 (0 byte read). Received data is not to be considered valid unless a complete record set is read successfully, including the closing 0-Byte read. You are advised to always read the complete set into a user space buffer before processing the data.

When designing a buffer, allow for record sizes up to the size of the entire monitor DCSS, or use dynamic memory allocation. The size of the monitor DCSS will be printed into syslog after loading the module. You can also use the (Class E privileged) CP command Q NSS MAP to list all available segments and information about them (see “Assuring that the DCSS is addressable for your Linux instance” on page 387).

Error conditions are indicated by returning a negative value for the number of bytes read. For an error condition, the errno variable can be:

**EIO**  
Reply failed. All data that was read since the last successful read with 0 size is not valid. Data is missing. The application must decide whether to continue reading subsequent data or to exit.

**EFAULT**  
Copy to user failed. All data that was read since the last successful read with 0 size is not valid. Data is missing. The application must decide whether to continue reading subsequent data or to exit.

**EAGAIN**  
Occurs on a non-blocking read if there is no data available at the moment. No data is missing or damaged. Retry or use polling for non-blocking reads.

**EOVERFLOW**  
The message limit is reached. The data that was read since the last successful read with 0 size is valid, but subsequent records might be missing. The application must decide whether to continue reading subsequent data or to exit.
Chapter 33. z/VM recording device driver

The z/VM recording device driver enables Linux on z/VM to read from the CP recording services and, thus, act as a z/VM wide control point.

The z/VM recording device driver uses the z/VM CP RECORDING command to collect records and IUCV to transmit them to the Linux instance.

For general information about CP recording system services, see z/VM CP Programming Services, SC24-6179.

Features

With the z/VM recording device driver, you can read from several CP services and collect records.

In particular, the z/VM recording device driver supports:

- Reading records from the CP error logging service, *LOGREC.
- Reading records from the CP accounting service, *ACCOUNT.
- Reading records from the CP diagnostic service, *SYMPTOM.
- Automatic and explicit record collection (see "Starting and stopping record collection" on page 393).

What you should know about the z/VM recording device driver

You can read records from different recording services, one record at a time.

The z/VM recording device driver is a character device driver that is grouped under the IUCV category of device drivers (see "Device categories" on page 7). There is one device for each recording service. The devices are created for you when the z/VM recording device driver module is loaded.

z/VM recording device nodes

Each recording service has a name that corresponds to the name of the service.

Table 50 summarizes the names:

<table>
<thead>
<tr>
<th>z/VM recording service</th>
<th>Standard device name</th>
</tr>
</thead>
<tbody>
<tr>
<td>*LOGREC</td>
<td>logrec</td>
</tr>
<tr>
<td>*ACCOUNT</td>
<td>account</td>
</tr>
<tr>
<td>*SYMPTOM</td>
<td>symptom</td>
</tr>
</tbody>
</table>

About records

Records for different services are different in details, but follow the same overall structure.

The read function returns one record at a time. If there is no record, the read function waits until a record becomes available.
Each record begins with a 4-byte field that contains the length of the remaining record. The remaining record contains the binary z/VM data followed by the four bytes X'454f5200' to mark the end of the record. These bytes build the zero-terminated ASCII string “EOR”, which is useful as an eye catcher.

![Record structure diagram]

Figure 76. Record structure

Figure 76 illustrates the structure of a complete record as returned by the device. If the buffer assigned to the read function is smaller than the overall record size, multiple reads are required to obtain the complete record.

The format of the z/VM data (*LOGREC) depends on the record type that is described in the common header for error records HDRREC.

For more information about the z/VM record layout, see the CMS and CP Data Areas and Control Blocks documentation at www.ibm.com/vm/pubs/ctlblk.html

**Setting up the z/VM recording device driver**

Before you can collect records, you must authorize your z/VM guest virtual machine and load the device driver module.

**Procedure**

1. Authorize the z/VM guest virtual machine on which your Linux instance runs to:
   - Use the z/VM CP RECORDING command.
   - Connect to the IUCV services to be used: one or more of *LOGREC, *ACCOUNT, and *SYMPTOM.

2. Load the z/VM recording device driver.
   You must load the z/VM recording device driver module before you can work with z/VM recording devices. Load the vmlogrdr module with the `modprobe` command to ensure that any other required modules are loaded in the correct order:

   ```
   # modprobe vmlogrdr
   ```

   There are no module parameters for the z/VM recording device driver.

**Working with z/VM recording devices**

Typical tasks that you perform with z/VM recording devices include starting and stopping record collection, purging records, and opening and closing devices.

- "Starting and stopping record collection" on page 393
- "Purging existing records" on page 394
- "Querying the z/VM recording status" on page 394
Starting and stopping record collection

By default, record collection for a particular z/VM recording service begins when the corresponding device is opened and stops when the device is closed.

About this task

You can use a device's autorecording attribute to be able to open and close a device without also starting or stopping record collection. You can use a device's recording attribute to start and stop record collection regardless of whether the device is opened or not.

You cannot start record collection if a device is open and records already exist. Before you can start record collection for an open device, you must read or purge any existing records for this device (see “Purging existing records” on page 394).

Procedure

To be able to open a device without starting record collection and to close a device without stopping record collection, write 0 to the device's autorecording attribute. To restore the automatic starting and stopping of record collection, write 1 to the device’s autorecording attribute. Issue a command of this form:

```
# echo <flag> > /sys/bus/iucv/drivers/vmlogrdr/<device>/autorecording
```

where `<flag>` is either 0 or 1, and `<device>` is one of: logrec, symptom, or account. To explicitly turn on record collection write 1 to the device’s recording attribute. To explicitly turn off record collection write 0 to the device’s recording attribute. Issue a command of this form:

```
# echo <flag> > /sys/bus/iucv/drivers/vmlogrdr/<device>/recording
```

where `<flag>` is either 0 or 1, and `<device>` is one of: logrec, symptom, or account. You can read both the autorecording and the recording attribute to find the current settings.

Examples

- In this example, first the current setting of the autorecording attribute of the logrec device is checked, then automatic recording is turned off:

```
# cat /sys/bus/iucv/drivers/vmlogrdr/logrec/autorecording
1
# echo 0 > /sys/bus/iucv/drivers/vmlogrdr/logrec/autorecording
```

- In this example record collection is started explicitly and later stopped for the account device:

```
# echo 1 > /sys/bus/iucv/drivers/vmlogrdr/account/recording
...
# echo 0 > /sys/bus/iucv/drivers/vmlogrdr/account/recording
```

To confirm whether recording is on or off, read the recording_status attribute as described in “Querying the z/VM recording status” on page 394.
Purging existing records

By default, existing records for a particular z/VM recording service are purged automatically when the corresponding device is opened or closed.

About this task

You can use a device's autopurge attribute to prevent records from being purged when a device is opened or closed. You can use a device's purge attribute to purge records for a particular device at any time without having to open or close the device.

Procedure

To be able to open or close a device without purging existing records write 0 to the device's autopurge attribute. To restore automatic purging of existing records, write 1 to the device's autopurge attribute. You can read the autopurge attribute to find the current setting. Issue a command of this form:

```
# echo <flag> > /sys/bus/iucv/drivers/vmlogrdr/<device>/autopurge
```

where <flag> is either 0 or 1, and <device> is one of: logrec, symptom, or account.

To purge existing records for a particular device without opening or closing the device, write 1 to the device's purge attribute. Issue a command of this form:

```
# echo 1 > /sys/bus/iucv/drivers/vmlogrdr/<device>/purge
```

where <device> is one of: logrec, symptom, or account.

Examples

- In this example, the setting of the autopurge attribute for the logrec device is checked first, then automatic purging is switched off:

  ```
  # cat /sys/bus/iucv/drivers/vmlogrdr/logrec/autopurge
  1
  # echo 0 > /sys/bus/iucv/drivers/vmlogrdr/logrec/autopurge
  ```

- In this example, the existing records for the symptom device are purged:

  ```
  # echo 1 > /sys/bus/iucv/drivers/vmlogrdr/symptom/purge
  ```

Querying the z/VM recording status

Use the recording_status attribute of the z/VM recording device driver representation in sysfs to query the z/VM recording status.

Example

This example runs the z/VM CP command QUERY RECORDING and returns the complete output of that command. This list does not necessarily have an entry for all three services and there might also be entries for other guests.

```
# cat /sys/bus/iucv/drivers/vmlogrdr/recording_status
```

This command results in output similar to the following example:
where the lines represent:

- The service
- The recording status
- The number of queued records
- The number of records that result in a message to the operator
- The guest that is or was connected to that service and the status of that connection

A detailed description of the QUERY RECORDING command can be found in the *z/VM CP Commands and Utilities Reference, SC24-6175.

### Opening and closing devices

You can open, read, and release the device. You cannot open the device multiple times. Each time the device is opened it must be released before it can be opened again.

#### About this task

You can use a device’s autorecord attribute (see “Starting and stopping record collection” on page 393) to enable automatic record collection while a device is open.

You can use a device’s autopurge attribute (see “Purging existing records” on page 394) to enable automatic purging of existing records when a device is opened and closed.

### Scenario: Connecting to the *ACCOUNT service

A typical sequence of tasks is autorecording, turning autorecording off, purging records, and starting recording.

#### Procedure

1. Query the status of z/VM recording. As root, issue the following command:

   ```
   # cat /sys/bus/iucv/drivers/vmlogrdr/recording_status
   ```

   The results depend on the system, and look similar to the following example:

   ~~~
   RECORDING COUNT LMT USERID COMMUNICATION
   EREP ON 00000000 002 EREP ACTIVE
   ACCOUNT ON 00001812 020 DISKACNT INACTIVE
   SYMPTOM ON 00000000 002 OPERSYMP ACTIVE
   ACCOUNT OFF 00000000 020 LINUX31 INACTIVE
   ~~~

2. Open /dev/account with an appropriate application. This action connects the guest to the *ACCOUNT service and starts recording. The entry for *ACCOUNT on guest LINUX31 changes to ACTIVE and ON:
3. Switch autopurge and autorecord off:

```bash
# echo 0 > /sys/bus/iucv/drivers/vmlogrdr/account/autopurge

# echo 0 > /sys/bus/iucv/drivers/vmlogrdr/account/autorecording
```

4. Close the device by ending the application that reads from it and check the recording status. While the connection is INACTIVE, RECORDING is still ON:

```bash
# cat /sys/bus/iucv/drivers/vmlogrdr/recording_status
```

5. The next status check shows that some event created records on the *ACCOUNT queue:

```bash
# cat /sys/bus/iucv/drivers/vmlogrdr/recording_status
```

6. Switch recording off:

```bash
# echo 0 > /sys/bus/iucv/drivers/vmlogrdr/account/recording

# cat /sys/bus/iucv/drivers/vmlogrdr/recording_status
```

7. Try to switch it on again, and check whether it worked by checking the recording status:

```bash
# echo 1 > /sys/bus/iucv/drivers/vmlogrdr/account/recording

# cat /sys/bus/iucv/drivers/vmlogrdr/recording_status
```

Recording did not start, in the message logs you might find a message:

`vmlogrdr: recording response: HCPCR08087I Records are queued for user LINUX31 on the ACCOUNT recording queue and must be purged or retrieved before recording can be turned on.`
This kernel message has priority ‘debug’ so it might not be written to any of your log files.

8. Now remove all the records on your *ACCOUNT queue either by starting an application that reads them from /dev/account or by explicitly purging them:

   # echo 1 > /sys/bus/iucv/drivers/vmlogrdr/account/purge

   # cat /sys/bus/iucv/drivers/vmlogrdr/recording_status
   RECORDING  COUNT  LMT  USERID  COMMUNICATION
   EREP ON    00000000 002  EREP  ACTIVE
   ACCOUNT ON 00001821 020  DISKACNT INACTIVE
   SYMPTOM ON 00000000 002  OPERSYMP ACTIVE
   ACCOUNT OFF 00000000 020  LINUX31 INACTIVE

9. Now start recording and check status again:

   # echo 1 > /sys/bus/iucv/drivers/vmlogrdr/account/recording

   # cat /sys/bus/iucv/drivers/vmlogrdr/recording_status
   RECORDING  COUNT  LMT  USERID  COMMUNICATION
   EREP ON    00000000 002  EREP  ACTIVE
   ACCOUNT ON 00001821 020  DISKACNT INACTIVE
   SYMPTOM ON 00000000 002  OPERSYMP ACTIVE
   ACCOUNT ON 00000000 020  LINUX31 INACTIVE
Chapter 34. z/VM unit record device driver

The z/VM unit record device driver provides Linux on z/VM with access to virtual unit record devices. Unit record devices comprise punch card readers, card punches, and line printers.

Linux access is limited to virtual unit record devices with default device types (2540 for reader and punch, 1403 for printer).

To write Linux files to the virtual punch or printer (that is, to the corresponding spool file queues) or to receive z/VM reader files (for example CONSOLE files) to Linux files, use the vmur command that is part of the s390utils RPM (see “vmur - Work with z/VM spool file queues” on page 631).

What you should know about the z/VM unit record device driver

The z/VM unit record device driver is compiled as a separate module, vmur.

To load the vmur module automatically at boot time, see the section on persistent module loading in Red Hat Enterprise Linux 7.5 System Administrator’s Guide

When the vmur module is loaded, it registers a character device. The following device nodes are created for a unit record device when it is set online:

- Reader: /dev/vmrdr-0.0.<device_number>
- Punch: /dev/vmpun-0.0.<device_number>
- Printer: /dev/vmprt-0.0.<device_number>

Working with z/VM unit record devices

After loading the vmur module, the required virtual unit record devices must be set online.

Procedure

Set the virtual unit record devices online. For example, to set the devices with device bus-IDs 0.0.000c, 0.0.000d, and 0.0.000e online, issue:

```
# chccwdev -e 0.0.000c-0.0.000e
```

What to do next

You can now use the vmur command to work with the devices (“vmur - Work with z/VM spool file queues” on page 631).

If you want to unload the vmur module, close all unit record device nodes. Attempting to unload the module while a device node is open results in error message Module vmur is in use. You can unload the vmur module, for example, by issuing modprobe -r.
Serialization is implemented per device; only one process can open a particular device node at any one time.
Chapter 35. z/VM DCSS device driver

The z/VM discontiguous saved segments (DCSS) device driver provides disk-like fixed block access to z/VM discontiguous saved segments.

In particular, the DCSS device driver facilitates implementing a read-write RAM disk that can be shared among multiple Linux instances that run as guests of the same z/VM system. For example, such a RAM disk can provide a shared file system.

For information about DCSS, see z/VM Saved Segments Planning and Administration, SC24-6229

What you should know about DCSS

The DCSS device names and nodes adhere to a naming scheme. There are different modes and options for mounting a DCSS.

Important: DCSSs occupy spool space. Be sure that you have enough spool space available (multiple times the DCSS size).

DCSS naming scheme

The standard device names are of the form dcssblk<n>, where <n> is the corresponding minor number.

The first DCSS device that is added is assigned the name dcssblk0, the second dcssblk1, and so on. When a DCSS device is removed, its device name and corresponding minor number are free and can be reassigned. A DCSS device that is added always receives the lowest free minor number.

DCSS device nodes

User space programs access DCSS devices by device nodes. Red Hat Enterprise Linux 7.5 creates standard DCSS device nodes for you.

Standard DCSS device nodes have the form /dev/<device_name>, for example:

```
/dev/dcssblk0
/dev/dcssblk1
...```

Accessing a DCSS in exclusive-writable mode

You must access a DCSS in exclusive-writable mode, for example, to create or update the DCSS.

To access a DCSS in exclusive-writable mode at least one of the following conditions must apply:

- The DCSS fits below the maximum definable address space size of the z/VM guest virtual machine.

For large read-only DCSS, you can use suitable guest sizes to restrict exclusive-writable access to a specific z/VM guest virtual machine with a sufficient maximum definable address space size.
The z/VM user directory entry for the z/VM guest virtual machine includes a NAMESAVE statement for the DCSS. See z/VM CP Planning and Administration, SC24-6178 for more information about the NAMESAVE statement.

The DCSS has been defined with the LOADNSHR operand. See z/VM CP Commands and Utilities Reference, SC24-6175 for information about the LOADNSHR operand.

See "DCSS options" about saving DCSSs with the LOADNSHR operand or with other optional properties.

**DCSS options**

The z/VM DCSS device driver always saves DCSSs with default properties. Any previously defined options are removed.

For example, a DCSS that was defined with the LOADNSHR operand loses this property when it is saved with the z/VM DCSS device driver.

To save a DCSS with optional properties, you must unmount the DCSS device, then use the CP DEFSEG and SAVESEG commands to save the DCSS. See "Workaround for saving DCSSs with optional properties" on page 408 for an example.

See z/VM CP Commands and Utilities Reference, SC24-6175 for information about DCSS options.

**Setting up the DCSS device driver**

Before you can load and use DCSSs, you must load the DCSS block device driver. Use the segments module parameter to load one or more DCSSs when the DCSS device driver is loaded.

---

**DCSS module parameter syntax**

```
modprobe dcssblk segments=<dcss>(local)
```

- `<dcss>` specifies the name of a DCSS as defined on the z/VM hypervisor. The specification for `<dcss>` is converted from ASCII to uppercase EBCDIC.

- The colon (:) separates DCSSs within a set of DCSSs to be mapped to a single DCSS device. You can map a set of DCSSs to a single DCSS device if the DCSSs in the set form a contiguous memory space.

  You can specify the DCSSs in any order. The name of the first DCSS you specify is used to represent the device under `/sys/devices/dcssblk`.

- `(local)` sets the access mode to exclusive-writable after the DCSS or set of DCSSs are loaded.
the comma (,) separates DCSS devices.

**Examples**

The following command loads the DCSS device driver and three DCSSs: DCSS1, DCSS2, and DCSS3. DCSS2 is accessed in exclusive-writable mode.

```bash
# modprobe dcssblk segments="dcss1,dcss2(local),dcss3"
```

The following command loads the DCSS device driver and four DCSSs: DCSS4, DCSS5, DCSS6, and DCSS7. The device driver creates two DCSS devices. One device maps to DCSS4 and the other maps to the combined storage space of DCSS5, DCSS6, and DCSS7 as a single device.

```bash
# modprobe dcssblk segments="dcss4,dcss5:dcss6:dcss7"
```

**Avoiding overlaps with your guest storage**

Ensure that your DCSSs do not overlap with the memory of your z/VM guest virtual machine (guest storage).

**About this task**

To find the start and end addresses of the DCSSs, enter the following CP command; this command requires privilege class E:

```bash
# cp q nss map
```

the output gives you the start and end addresses of all defined DCSSs in units of 4-kilobyte pages:

```
00: FILE FILENAME FILETYPE MINSIZE BEGPAG ENDPAG TYPE CL #USERS PARMREGS VMGROUP
...
00: 0011 MONDCSS CPDCSS N/A 09000 097FF SC R 00003 N/A N/A
...
```

If all DCSSs that you intend to access are located above the guest storage, you do not need to take any action.

**Procedure**

If any DCSS that you intend to access with your guest machine overlaps with the guest storage, redefine the guest storage. Define two or more discontiguous storage extents such that the storage gap with the lowest address range covers the address ranges of all your DCSSs.

**Note:**
- You cannot place a DCSS into a storage gap other than the storage gap with the lowest address range.
- A z/VM guest that was defined with one or more storage gaps cannot access a DCSS above the guest storage.
From a CMS session, use the DEF STORE command to define your guest storage as discontiguous storage extents. Ensure that the storage gap between the extents covers all your DCSSs’ address ranges. Issue a command of this form:

```
DEF STOR CONFIG 0.<storage_gap_begin> <storage_gap_end>.<storage above gap>
```

where:

<storage_gap_begin>

is the lower limit of the storage gap. This limit must be at or below the lowest address of the DCSS with the lowest address range.

Because the lower address ranges are needed for memory management functions, make the lower limit at least 128 MB. The lower limit for the DCSS increases with the total memory size. Although 128 MB is not an exact value, it is an approximation that is sufficient for most cases.

<storage_gap_end>

is the upper limit of the storage gap. The upper limit must be above the upper limit of the DCSS with the highest address range.

<storage above gap>

is the amount of storage above the storage gap. The total guest storage is <storage_gap_begin> + <storage above gap>.

All values can be suffixed with M to provide the values in megabyte. See z/VM CP Commands and Utilities Reference, SC24-6175 for more information about the DEF STORE command.

**Example**

To make a DCSS that starts at 144 MB and ends at 152 MB accessible to a z/VM guest with 512 MB guest storage:

```
DEF STORE CONFIG 0.140M 160M.372M
```

This specification is one example of how a suitable storage gap can be defined. In this example, the storage gap covers 140 - 160 MB and, thus, the entire DCSS range. The total guest storage is 140 MB + 372 MB = 512 MB.

**Working with DCSS devices**

Typical tasks for working with DCSS devices include mapping DCSS representations in z/VM and Linux, adding and removing DCSSs, and accessing and updating DCSS contents.

- “Adding a DCSS device” on page 405
- “Listing the DCSSs that map to a particular device” on page 405
- “Finding the minor number for a DCSS device” on page 406
- “Setting the access mode” on page 406
- “Saving updates to a DCSS or set of DCSSs” on page 407
- “Workaround for saving DCSSs with optional properties” on page 408
- “Removing a DCSS device” on page 409
Adding a DCSS device

Storage gaps or overlapping storage ranges can prevent you from adding a DCSS.

Before you begin

- You must have set up one or more DCSSs on z/VM and know their names on z/VM.
- If you use the watchdog device driver, turn off the watchdog before adding a DCSS device. Adding a DCSS device can result in a watchdog timeout if the watchdog is active.
- You cannot concurrently access overlapping DCSSs.
- You cannot access a DCSS that overlaps with your z/VM guest virtual storage (see “Avoiding overlaps with your guest storage” on page 403).
- On z/VM guest virtual machines with one or more storage gaps, you cannot add a DCSS that is above the guest storage.
- On z/VM guest virtual machines with multiple storage gaps, you cannot add a DCSS unless it fits in the storage gap with the lowest address range.

Procedure

To add a DCSS device, enter a command of this form:

```
# echo <dcss-list> > /sys/devices/dcssblk/add
```

<dcss-list>
the name, as defined on z/VM, of a single DCSS or a colon (:) separated list of names of DCSSs to be mapped to a single DCSS device. You can map a set of DCSSs to a single DCSS device if the DCSSs in the set form a contiguous memory space. You can specify the DCSSs in any order. The name of the first DCSS you specify is used to represent the device under /sys/devices/dcssblk.

Examples

To add a DCSS called “MYDCSS” enter:

```
# echo MYDCSS > /sys/devices/dcssblk/add
```

To add three contiguous DCSSs “MYDCSS1”, “MYDCSS2”, and “MYDCSS3” as a single device, enter:

```
# echo MYDCSS2:MYDCSS1:MYDCSS3 > /sys/devices/dcssblk/add
```

In sysfs, the resulting device is represented as /sys/devices/dcssblk/MYDCSS2.

Listing the DCSSs that map to a particular device

Read the seglist sysfs attribute to find out how DCSS devices in Linux map to DCSSs as defined in z/VM.

Procedure

To list the DCSSs that map to a DCSS device, issue a command of this form:

```
# cat /sys/devices/dcssblk/<dcss-name>/seglist
```
where `<dcss-name>` is the DCSS name that represents the DCSS device.

**Examples**

In this example, DCSS device `MYDCSS` maps to a single DCSS, “MYDCSS”.

```
# cat /sys/devices/dcssblk/MYDCSS/seglist
MYDCSS
```

In this example, DCSS device `MYDCSS2` maps to three contiguous DCSSs, “MYDCSS1”, “MYDCSS2”, and “MYDCSS3”.

```
# cat /sys/devices/dcssblk/MYDCSS2/seglist
MYDCSS2
MYDCSS1
MYDCSS3
```

**Finding the minor number for a DCSS device**

When you add a DCSS device, a minor number is assigned to it.

**About this task**

Unless you use dynamically created device nodes as provided by udev, you might need to know the minor device number that was assigned to the DCSS (see “DCSS naming scheme” on page 401).

When you add a DCSS device, a directory of this form is created in sysfs:

```
/sys/devices/dcssblk/<dcss-name>
```

where `<dcss-name>` is the DCSS name that represents the DCSS device.

This directory contains a symbolic link, block, that helps you to find out the device name and minor number. The link is of the form `../../../block/dcssblk<n>`, where `dcssblk<n>` is the device name and `<n>` is the minor number.

**Example**

To find out the minor number that is assigned to a DCSS device that is represented by the directory `/sys/devices/dcssblk/MYDCSS` issue:

```
# readlink /sys/devices/dcssblk/MYDCSS/block
../../../block/dcssblk0
```

In the example, the assigned minor number is 0.

**Setting the access mode**

You might want to access the DCSS device with write access to change the content of the DCSS or set of DCSSs that map to the device.

**About this task**

There are two possible write access modes to the DCSS device:
In the shared mode, changes to DCSSs are immediately visible to all z/VM guests that access them. Shared is the default.

Note: Writing to a shared DCSS device bears the same risks as writing to a shared disk.

In the exclusive-writable mode you write to private copies of DCSSs. A private copy is writable, even if the original DCSS is read-only. Changes that you make to a private copy are invisible to other guests until you save the changes (see “Saving updates to a DCSS or set of DCSSs”).

After saving the changes to a DCSS, all guests that open the DCSS access the changed copy. z/VM retains a copy of the original DCSS for those guests that continue accessing it, until the last guest stops using it.

To access a DCSS in the exclusive-writable mode the maximum definable storage size of your z/VM virtual machine must be above the upper limit of the DCSS. Alternatively, suitable authorizations must be in place (see “Accessing a DCSS in exclusive-writable mode” on page 401).

For either access mode the changes are volatile until they are saved (see “Saving updates to a DCSS or set of DCSSs”).

Procedure

Set the access mode before you open the DCSS device. To set the access mode to exclusive-writable, set the DCSS device’s shared attribute to 0. To reset the access mode to shared set the DCSS device’s shared attribute to 1.

Issue a command of this form:

```
# echo <flag> > /sys/devices/dcssblk/<dcss-name>/shared
```

where `<dcss-name>` is the DCSS name that represents the DCSS device. You can read the shared attribute to find out the current access mode.

Example

To find out the current access mode of a DCSS device represented by the DCSS name “MYDCSS”:

```
# cat /sys/devices/dcssblk/MYDCSS/shared
1
```

1 means that the current access mode is shared. To set the access mode to exclusive-writable issue:

```
# echo 0 > /sys/devices/dcssblk/MYDCSS/shared
```

Saving updates to a DCSS or set of DCSSs

Use the `save` sysfs attribute to save DCSSs that were defined without optional properties.
Before you begin

- Saving a DCSS as described in this section results in a default DCSS, without optional properties. For DCSSs that were defined with options (see "DCSS options" on page 402), see "Workaround for saving DCSSs with optional properties."
- If you use the watchdog device driver, turn off the watchdog before saving updates to DCSSs. Saving updates to DCSSs can result in a watchdog timeout if the watchdog is active.
- Do not place save requests before you have accessed the DCSS device.

Procedure

To place a request for saving changes permanently on the spool disk write 1 to the DCSS device's save attribute. If a set of DCSSs has been mapped to the DCSS device, the save request applies to all DCSSs in the set.

Issue a command of this form:

```
# echo 1 > /sys/devices/dcssblk/<dcss-name>/save
```

where <dcss-name> is the DCSS name that represents the DCSS device. Saving is delayed until you close the device.

You can check if a save request is waiting to be performed by reading the contents of the save attribute.

You can cancel a save request by writing 0 to the save attribute.

Example

To check if a save request exists for a DCSS device that is represented by the DCSS name “MYDCSS”:

```
# cat /sys/devices/dcssblk/MYDCSS/save
```

The 0 means that no save request exists. To place a save request issue:

```
# echo 1 > /sys/devices/dcssblk/MYDCSS/save
```

To purge an existing save request issue:

```
# echo 0 > /sys/devices/dcssblk/MYDCSS/save
```

Workaround for saving DCSSs with optional properties

If you need a DCSS that is defined with special options, you must use a workaround to save the DCSSs.

Before you begin

Important: This section applies to DCSSs with special options only. The workaround in this section is error-prone and requires utmost care. Erroneous parameter values for the described CP commands can render a DCSS unusable. Use this workaround only if you really need a DCSS with special options.
**Procedure**

Perform the following steps to save a DCSS with optional properties:

1. Unmount the DCSS.

   **Example:** Enter this command to unmount a DCSS with the device node `/dev/dcssblk0`:
   ```bash
   # umount /dev/dcssblk0
   ```

2. Use the CP DEFSEG command to newly define the DCSS with the required properties.

   **Example:** Enter this command to newly define a DCSS, mydcss, with the range `80000-9ffff`, segment type `sr`, and the `loadnshr` operand:
   ```bash
   # vmcp defseg mydcss 80000-9ffff sr loadnshr
   ```

   **Note:** If your DCSS device maps to multiple DCSSs as defined to z/VM, you must perform this step for each DCSS. Be sure to specify the command correctly with the correct address ranges and segment types. Incorrect specifications can render the DCSS unusable.

3. Use the CP SAVESEG command to save the DCSS.

   **Example:** Enter this command to save a DCSS `mydcss`:
   ```bash
   # vmcp saveseg mydcss
   ```

   **Note:** If your DCSS device maps to multiple DCSSs as defined to z/VM, you must perform this step for each DCSS. Omitting this step for individual DCSSs can render the DCSS device unusable.

**Reference**

See *z/VM CP Commands and Utilities Reference*, SC24-6175 for details about the DEFSEG and SAVESEG CP commands.

**Removing a DCSS device**

Use the `remove sysfs` attribute to remove a DCSS from Linux.

**Before you begin**

A DCSS device can only be removed when it is not in use.

**Procedure**

You can remove the DCSS or set of DCSSs that are represented by a DCSS device from your Linux system by issuing a command of this form:

```bash
# echo <dcss-name> > /sys/devices/dcssblk/remove
```

where `<dcss-name>` is the DCSS name that represents the DCSS device.
Example

To remove a DCSS device that is represented by the DCSS name “MYDCSS” issue:

```
# echo MYDCSS > /sys/devices/dcssblk/remove
```

What to do next

If you have created your own device nodes, you can keep the nodes for reuse. Be aware that the major number of the device might change when you unload and reload the DCSS device driver. When the major number of your device has changed, existing nodes become unusable.

Scenario: Changing the contents of a DCSS

Before you can change the contents of a DCSS, you must add the DCSS to Linux, access it in a writable mode, and mount the file system on it.

About this task

The scenario that follows is based on these assumptions:

- The Linux instance runs as a z/VM guest with class E user privileges.
- A DCSS is set up and can be accessed in exclusive-writable mode by the Linux instance.
- The DCSS does not overlap with the guest's main storage.
- There is only a single DCSS named “MYDCSS”.
- The DCSS block device driver is set up and ready to be used.

The description in this scenario can readily be extended to changing the content of a set of DCSSs that form a contiguous memory space. The only change to the procedure would be mapping the DCSSs in the set to a single DCSS device in step 1. The assumptions about the set of DCSSs would be:

- The contiguous memory space that is formed by the set does not overlap with the guest storage.
- Only the DCSSs in the set are added to the Linux instance.

Procedure

Perform the following steps to change the contents of a DCSS:

1. Add the DCSS to the block device driver.

```
# echo MYDCSS > /sys/devices/dcssblk/add
```

2. Ensure that there is a device node for the DCSS block device. If it is not created for you, for example by udev, create it yourself.

   a. Find out the major number that is used for DCSS block devices. Read `/proc/devices`:

```
# cat /proc/devices
```

   ... Block devices ...
   ... 254 dcssblk ...


The major number in the example is 254.

b. Find out the minor number that is used for MYDCSS. If MYDCSS is the first DCSS to be added, the minor number is 0. To be sure, you can read a symbolic link that is created when the DCSS is added.

```bash
# readlink /sys/devices/dcssblk/MYDCSS/block
../../../../../block/dcssblk0
```

The trailing 0 in the standard device name dcssblk0 indicates that the minor number is, indeed, 0.

c. Create the node with the `mknod` command:

```bash
# mknod /dev/dcssblk0 b 254 0
```

3. Set the access mode to exclusive-write.

```bash
# echo 0 > /sys/devices/dcssblk/MYDCSS/shared
```

4. Mount the file system in the DCSS on a spare mount point.

```bash
# mount /dev/dcssblk0 /mnt
```

5. Update the data in the DCSS.

6. Create a save request to save the changes.

```bash
# echo 1 > /sys/devices/dcssblk/MYDCSS/save
```

7. Unmount the file system.

```bash
# umount /mnt
```

The changes to the DCSS are now saved. When the last z/VM guest stops accessing the old version of the DCSS, the old version is discarded. Each guest that opens the DCSS accesses the updated copy.

8. Remove the device.

```bash
# echo MYDCSS > /sys/devices/dcssblk/remove
```

9. Optional: If you have created your own device node, you can clean it up.

```bash
# rm -f /dev/dcssblk0
```
Chapter 36. z/VM CP interface device driver

Using the z/VM CP interface device driver (vmcp), you can send control program (CP) commands to the z/VM hypervisor and display the response.

The vmcp device driver works only for Linux on z/VM.

What you should know about the z/VM CP interface

The z/VM CP interface device driver (vmcp) uses the CP diagnose X'08' to send commands to CP and to receive responses. The behavior is similar but not identical to #CP on a 3270 or 3215 console.

Using the z/VM CP interface

There are two ways of using the z/VM CP interface device driver:

- Through the /dev/vmcp device node
- Through a user space tool (see "vmcp - Send CP commands to the z/VM hypervisor" on page 629)

You must load the vmcp module before you can use vmcp. If your Linux guest runs under z/VM, you can configure the startup scripts to load the vmcp kernel module automatically during boot. See the section on persistent module loading in Red Hat Enterprise Linux 7.5 System Administrator's Guide for information about how to do this.

The vmcp device driver works only for Linux on z/VM and cannot be loaded if the Linux system runs in an LPAR.

Differences between vmcp and a 3270 or 3215 console

Most CP commands behave identically with vmcp and on a 3270 or 3215 console. However, some commands show a different behavior:

- Diagnose X'08' (see z/VM CP Programming Services, SC24-6179) requires you to specify a response buffer with the command. Because the response size is not known in advance, the default response buffer of vmcp might be too small and the response truncated.
- On a 3270 or 3215 console, the CP command is executed on virtual CPU 0. The vmcp device driver uses the CPU that is scheduled by the Linux kernel. For CP commands that depend on the CPU number (like trace), specify the CPU, for example: cpu 3 trace count.
- Some CP commands do not return specific error or status messages through diagnose X'08'. These messages are only returned on a 3270 or 3215 console. For example, the command vmcp link user1 1234 123 mw might return the message DASD 123 LINKED R/W in a 3270 or 3215 console. This message is not displayed if the CP command is issued with vmcp. For details, see the z/VM help system or z/VM CP Commands and Utilities Reference, SC24-6175.
Using the device node

You can send a command to z/VM CP by writing to the vmcp device node.

Observe the following rules for writing to the device node:

- Omit the newline character at the end of the command string. For example, use `echo -n` if you are writing directly from a terminal session.
- Write the command in the same case as required on z/VM.
- Escape characters that need escaping in the environment where you issue the command.

Example

The following command attaches a device to your z/VM guest virtual machine. The asterisk (*) is escaped to prevent the command shell from interpreting it.

```
# echo -n ATTACH 1234 \* > /dev/vmcp
```

Application programmers

You can also use the vmcp device node directly from an application by using open, write (to issue the command), read (to get the response), ioctl (to get and set status), and close. The following ioctls are supported:

<table>
<thead>
<tr>
<th>Name</th>
<th>Code definition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMCP_GETCODE</td>
<td>_IOR(0x10, 1, int)</td>
<td>Queries the return code of z/VM.</td>
</tr>
<tr>
<td>VMCP_SETBUF</td>
<td>_IOW(0x10, 2, int)</td>
<td>Sets the buffer size (the device driver has a default of 4 KB; vmcp calls this ioctl to set it to 8 KB instead).</td>
</tr>
<tr>
<td>VMCP_GETSIZE</td>
<td>_IOR(0x10, 3, int)</td>
<td>Queries the size of the response.</td>
</tr>
</tbody>
</table>
Chapter 37. z/VM special messages uevent support

The smsgiucv_app kernel device driver receives z/VM CP special messages (SMSP) and delivers these messages to user space as udev events (uevents).

The device driver receives only messages that start with APP. The generated uevents contain the message sender and content as environment variables (see Figure 77).

You can restrict the received special messages to a particular z/VM user ID. CP special messages are discarded if the specified sender does not match the sender of the CP special message.

Setting up the CP special message device driver

Configure the CP special message device driver when you load the device driver module.

The z/VM user ID does not require special authorizations to receive CP special messages. CP special messages can be issued from the local z/VM guest virtual machine or from other guest virtual machines. You can issue special messages from Linux or from a CMS or CP session.

Load the device driver module with the `modprobe` command.
smsgiucv_app syntax

```
modprobe smsgiucv_app sender=<user_ID>
```

Where:

`sender = <user_ID>`

permits CP special messages from the specified z/VM user ID only. CP special messages are discarded if the specified sender does not match the sender of the CP special message. If the `sender=` option is empty or not set, CP special messages are accepted from any z/VM user ID.

Lowercase characters are converted to uppercase.

To receive messages from several user IDs leave the `sender=` parameter empty, or do not specify it, and then filter with udev rules (see “Example udev rule” on page 418).

To load the smsgiucv_app module automatically at boot time, see the section on persistent module loading in Red Hat Enterprise Linux 7.5 System Administrator’s Guide.

---

**Working with CP special messages**

You might have to send, access, or respond to CP special messages.

- “Sending CP special messages”
- “Accessing CP special messages through uevent environment variables” on page 417
- “Writing udev rules for handling CP special messages” on page 417

### Sending CP special messages

Issue a CP SMSG command from a CP or CMS session or from Linux to send a CP special message.

**Procedure**

To send a CP special message to LXGUEST1 from Linux, enter a command of the following form:

```
# vmcp SMSG LXGUEST1 APP "<message text>
```

To send a CP special message to LXGUEST1, enter the following command from a CP or CMS session:

```
#CP SMSG LXGUEST1 APP <message text>
```

The special messages cause uevents to be generated. See “Writing udev rules for handling CP special messages” on page 417 for information about handling the uevents.
Accessing CP special messages through udev environment variables

A udev for a CP special message contains environment variables that you can use to access the message.

**SM SG_ID**
Specifies the message prefix. The SMSG_ID environment variable is always set to APP, which is the prefix that is assigned to the smsgiucv_app device driver.

**SM SG_SENDER**
Specifies the z/VM user ID that sent the CP special message.

Use SMSG_SENDER in udev rules for filtering the z/VM user ID if you want to accept CP special messages from different senders. All alphabetic characters in the z/VM user ID are uppercase characters.

**SM SG_TEXT**
Contains the message text of the CP special message. The APP prefix and leading white spaces are removed.

Writing udev rules for handling CP special messages

When using the CP special messages device driver, CP special messages trigger uevents.

**change events**
The smsgiucv_app device driver generates change uevents for each CP special message that is received.

For example, the special message:

```
#CP SMSG LXGUEST1 APP THIS IS A TEST MESSAGE
```

might trigger the following uevent:

```
UEVENT[1263487666.708881] change /devices/iucv/smsgiucv_app (iucv)
ACTION=change
DEVPATH=/devices/iucv/smsgiucv_app
SUBSYSTEM=iucv
SMSG_SENDER=MAINT
SMSG_ID=APP
SMSG_TEXT=THIS IS A TEST MESSAGE
DRIVER=SMSGIUCV
SEQNUM=1493
```

**add and remove events**
In addition to the change event for received CP special messages, generic add and remove events are generated when the module is loaded or unloaded, for example:

```
UEVENT[1263487583.511146] add /module/smsgiucv_app (module)
ACTION=add
DEVPATH=/module/smsgiucv_app
SUBSYSTEM=module
SEQNUM=1487
```

```
UEVENT[1263487583.514622] add /devices/iucv/smsgiucv_app (iucv)
ACTION=add
DEVPATH=/devices/iucv/smsgiucv_app
SUBSYSTEM=iucv
DRIVER=SMSGIUCV
SEQNUM=1488
```
With the information from the uevents, you can create custom udev rules to trigger actions that depend on the settings of the SMSG_* environment variables (see "Accessing CP special messages through uevent environment variables" on page 417).

For your udev rules, use the add and remove uevents to initialize and clean up resources. To handle CP special messages, write udev rules that match change uevents. For more information about writing udev rules, see the udev man page.

**Example udev rule**

The udev rules that process CP special messages identify particular messages and define one or more specific actions as a response.

The following example shows how to process CP special messages by using udev rules. The example contains rules for actions, one for all senders and one for the MAINT, OPERATOR, and LNXADM senders only.

The rules are contained in a block that matches uevents from the smsgiucv_app device driver. If there is no match, processing ends:

```
# Sample udev rules for processing CP special messages.
#
# DEVPATH="/smsgiucv_app", GOTO="smsgiucv_app_end"
# -------- Rules for CP messages go here --------
LABEL="smsgiucv_app_end"
```

The example first uses the \texttt{vmur} command to load the vmur kernel module. Then the z/VM virtual punch device, 000d, is activated.

```
# --- Initialization ---

# load vmur and set the virtual punch device online
SUBSYSTEM="module", ACTION="add", RUN="/sbin/modprobe --quiet vmur"
SUBSYSTEM="module", ACTION="add", RUN="/sbin/chccwdev -e 000d"
```

418 Device Drivers, Features, and Commands - Red Hat Enterprise Linux 7.5
The following rule accepts messages from all senders. The message text must match the string UNAME. If it does, the output of the `uname` command (the node name and kernel version of the Linux instance) is sent back to the sender.

```
# --- Rules for all senders ----
# UNAME: tell the sender which kernel is running
ACTION=="change", ENV{SMSG_TEXT}=="UNAME", \ 
   PROGRAM=="/bin/uname -n -r", \ 
   RUN="/sbin/vmcp smsg $env{SMSG_SENDER} 'result'"
```

In the following example block rules are defined to accept messages from certain senders only. If no sender matches, processing ends. The message text must match the string DMESG. If it does, the environment variable PATH is set and the output of the `dmesg` command is sent into the z/VM reader of the sender. The name of the spool file is LINUX DMESG.

```
# --- Special rules available for particular z/VM user IDs ---
ENV{SMSG_SENDER}!="MAINT|OPERATOR|LNXADM", GOTO="smsgiucv_app_end"

# DMESG: punch dmesg output to sender
ACTION=="change", ENV{SMSG_TEXT}=="DMESG", \ 
   ENV{PATH}="/bin:/sbin:/usr/bin:/usr/sbin", \ 
   RUN="/bin/sh -c 'dmesg |fold -s -w 74 |vmur punch -r -t LINUX.DMESG -u $env{SMSG_SENDER}'"
```
Chapter 38. Cooperative memory management

Cooperative memory management (CMM, or "cmm1") can reduce the memory that is available to an instance of Linux on z/VM.

CMM allocates pages to page pools that are not available to Linux. A diagnose code indicates to z/VM that the pages in the page pools are out of use. z/VM can then immediately reuse these pages for other z/VM guests.

To set up CMM, you must perform these tasks:
1. Load the cmm module.
2. Set up a resource management tool that controls the page pool. This tool can be the z/VM resource monitor (VMRM) or a third-party systems management tool.

This chapter describes how to set up CMM. For background information about CMM, see "Cooperative memory management background" on page 373.

You can also use the cpuplugd command to define rules for cmm behavior, see "Basic configuration file for memory control" on page 513.

For information about setting up the external resource manager, see the chapter on VMRM in z/VM Performance, SC24-6208.

Setting up cooperative memory management

Set up Linux on z/VM to participate in the cooperative memory management by loading the cooperative memory management support module, cmm.

Use the modprobe command to load the module. See the modprobe man page for command details.

cooperative memory management module parameter syntax

```
modprobe cmm sender=VMRMSVM
```

where <user_ID> specifies the z/VM guest virtual machine that is permitted to send messages to the module through the special messages interface. The default z/VM user ID is VMRMSVM, which is the default for the VMRM service machine.

To load the cmm module automatically at boot time, see the section on persistent module loading in Red Hat Enterprise Linux 7.5 System Administrator’s Guide.
Example

To load the cooperative memory management module and allow the z/VM guest virtual machine TESTID to send messages:

```
# modprobe cmm sender=TESTID
```

Working with cooperative memory management

After it has been set up, CMM works through the resource manager. No further actions are necessary. You might want to read the sizes of the page pools for diagnostic purposes.

To reduce the Linux memory size, CMM allocates pages to page pools that make the pages unusable to Linux. There are two such page pools, a static pool and a timed pool. You can use the procsfs interface to read the sizes of the page pools.

Reading the size of the static page pool

You can read the current size of the static page pool from procsfs.

Procedure

Issue this command:

```
# cat /proc/sys/vm/cmm_pages
```

Reading the size of the timed page pool

You can read the current size of the timed page pool from procsfs.

Procedure

Issue this command:

```
# cat /proc/sys/vm/cmm_timed_pages
```
Part 7. Security

Chapter 39. Generic cryptographic device driver .................................. 425
Features .................................................................................. 425
What you should know about the cryptographic
device driver .......................................................................... 427
Setting up the cryptographic device driver ......................... 430
Working with cryptographic devices .................................. 432
External programming interfaces .................................... 439

Chapter 40. Pseudorandom number generator device driver .................. 441
Setting up the pseudo-random number device
driver ................................................................................... 441
Working with the PRNG device driver ............................. 442

Chapter 41. True random-number generator device driver .................. 445
Setting up the TRNG device driver ................................ 445

Chapter 42. Protected key device driver .................................. 447
Loading the device driver module ..................................... 447
External programming interfaces .................................... 447

Chapter 43. Hardware-accelerated in-kernel cryptography ..................... 449
Hardware dependencies and restrictions .......................... 449
Loading the support modules ............................................. 450
Confirming hardware support for cryptographic
operations ............................................................................. 451

Chapter 44. Instruction execution protection .................................. 453
Setting up instruction execution protection ...................... 453
Controlling stack execution protection ............................. 453

These device drivers and features support security aspects of Red Hat Enterprise
Linux 7.5 for z Systems.

Newest version

You can find the newest version of this book at
www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_distlibs.html

Restrictions

For prerequisites and restrictions see the IBM Z architecture specific information in
the Red Hat Enterprise Linux 7.5 release notes at
https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux
Chapter 39. Generic cryptographic device driver

The generic cryptographic device driver (zcrypt) supports cryptographic coprocessor and accelerator hardware. Cryptographic coprocessors provide secure key cryptographic operations for the IBM Common Cryptographic Architecture (CCA). As of CEX4P, cryptographic coprocessors support Enterprise PKCS #11 (EP11).

Some cryptographic processing in Linux can be offloaded from the CPU and performed by dedicated CCA or EP 11 coprocessors or accelerators. Several of these CCA or EP 11 coprocessors and accelerators are available offering a range of features. The generic cryptographic device driver (zcrypt) is required to use any available cryptographic hardware for processor offload.

Features

The generic cryptographic device driver supports a range of hardware and software functions.

Supported cryptographic adapters

The cryptographic hardware feature might contain one or two cryptographic adapters. Each adapter can be configured either as a coprocessor or as an accelerator. CEX5 and CEX4 cryptographic adapters can also be configured as EP11 coprocessors.

- Crypto Express5S (EP11) Coprocessor (CEX5P)
- Crypto Express5S Accelerator (CEX5A)
- Crypto Express5S (CCA) Coprocessor (CEX5C)
- Crypto Express4S (EP11) Coprocessor (CEX4P)
- Crypto Express4S Accelerator (CEX4A)
- Crypto Express4S (CCA) Coprocessor (CEX4C)
- Crypto Express3 Accelerator (CEX3A)
- Crypto Express3 Coprocessor (CEX3C)

For information about setting up your cryptographic environment on Linux under z/VM, see Security on z/VM, SG24-7471 and Security for Linux on System z, SG24-7728.

Supported facilities

The cryptographic device driver supports several cryptographic accelerators as well as CCA and EP11 coprocessors.

Cryptographic accelerators support clear key cryptographic algorithms. In particular, they provide fast RSA encryption and decryption for key sizes 1024-bit, 2048-bit, and 4096-bit (CEX5A, CEC4A, and CEC3A only).

Cryptographic coprocessors act as a hardware security module (HSM) and provide secure key cryptographic operations for the IBM Common Cryptographic Architecture (CCA) and the Enterprise PKCS#11 feature (EP11).


Cryptographic coprocessors also provide clear key RSA operations for 1024-bit, 2048-bit, and 4096-bit keys, and a true random number generator. The EP11 coprocessor supports only secure key operations.

**Hardware and software prerequisites**

Support for the Crypto Express5S, Crypto Express4S, and Crypto Express3 features depends on the IBM Z hardware model.

Table 52 lists the support for the cryptographic adapters.

<table>
<thead>
<tr>
<th>Cryptographic adapters</th>
<th>Mainframe support</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEX6A, CEX6C, and CEX6P</td>
<td>z14</td>
</tr>
<tr>
<td>CEX5A, CEX5C, and CEX5P</td>
<td>z14, z13, and z13s</td>
</tr>
<tr>
<td>CEX4A, CEX4C, and CEX4P</td>
<td>zEC12 and zBC12</td>
</tr>
<tr>
<td>CEX3A and CEX3C</td>
<td>zEC12, zBC12, z196 and z114</td>
</tr>
</tbody>
</table>

Table 53 lists the required software by function.

<table>
<thead>
<tr>
<th>Software required</th>
<th>Function that is supported by the software</th>
</tr>
</thead>
<tbody>
<tr>
<td>The CCA library</td>
<td>For the secure key cryptographic functions on CEX5C, CEX4C, or CEX3C features. For information about CEX5C, CEX4C, and CEX3C adapter coexistence and how to use CCA functions, see <em>Secure Key Solution with the Common Cryptographic Architecture Application Programmer’s Guide</em>, SC33-8294. You can obtain it at <a href="http://www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_cca.html">www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_cca.html</a>.</td>
</tr>
<tr>
<td>APAR VM65942</td>
<td>To support the z14 and the CEX6S adapter on z/VM 6.3 and 6.4.</td>
</tr>
<tr>
<td>APAR VM65577</td>
<td>To support CEX5A, CEX5C, and CEX5P adapters on z/VM 6.3 and 6.2. Note that EP11 support requires a dedicated adapter.</td>
</tr>
</tbody>
</table>


**What you should know about the cryptographic device driver**

Your use of the cryptographic device driver and the cryptographic hardware might require additional software. There are special considerations for Linux on z/VM, for performance, and for specific cryptographic operations.

**Functions provided by the cryptographic device driver**

The functions that are provided by the cryptographic device driver depend on whether it finds a cryptographic accelerator, a CCA coprocessor, or an EP11 coprocessor.

For both accelerators and CCA coprocessors, the device driver provides
Rivest-Shamir-Adleman (RSA) encryption and RSA decryption functions using clear keys. RSA operations are supported in both the modulus-exponent and the Chinese-Remainder Theorem (CRT) variants for any key size in the range 57 - 4096 bit.

For CCA coprocessors, the device driver also provides a function to pass CCA requests to the cryptographic coprocessor and an access to the true random number generator of the CCA coprocessor.

For EP11 coprocessors, the device driver provides functionality to pass EP11 requests to the cryptographic coprocessor.

**Adapter discovery**

The cryptographic device driver provides two misc device nodes, one for cryptographic requests, and one for a device from which random numbers can be read.

Cryptographic adapters are detected automatically when the module is loaded. They are reprobed periodically, and following any hardware problem.

Upon detection of a cryptographic adapter, the device driver presents a Linux misc device, z90crypt, to user space. A user space process can open the misc device to submit cryptographic requests to the adapter through IOCTLs.

If at least one of the detected cryptographic adapters is a coprocessor, an additional misc device, hw rng, is created from which random numbers can be read.

You can set cryptographic adapters online or offline in the device driver. The cryptographic device driver ignores adapters that are configured offline even if the hardware is detected. The online or offline configuration is independent of the hardware configuration.

**Request processing**

Cryptographic adapters process requests asynchronously.

The device driver detects request completion either by standard polling, a special high-frequency polling thread, or by hardware interrupts. Hardware interrupt support is only available for Linux instances that run in an LPAR. If hardware interrupt support is available, the device driver does not use polling to detect request completion.
All requests to either of the two misc devices are routed to a cryptographic adapter using a crypto request scheduling function that, for each adapter, takes into account:

- The supported functions
- The number of pending requests
- A speed rating

The cryptographic device driver supports multiple cryptographic domains simultaneously, see “Cryptographic domains.”

The device driver uses at least one domain for all adapters. If none is given, the kernel selects a default domain. Alternatively, you can select the default domain using a module parameter (see “Module parameters” on page 430).

Cryptographic domains

Crypto Express hardware adapters, coprocessors or accelerators, are divided into multiple domains, also called cryptographic domains or AP domains.

Each domain acts as an independent cryptographic device with its own state, including its own master key. Two domains in the same Crypto Express adapter are completely isolated and cannot access each other’s states. The maximum number of domains depends on the mainframe model and is the same for all Crypto Express adapters in that mainframe. For example, a z13 supports up to 85 domains (with hexadecimal domain IDs 0000 to 005d).

Cryptographic devices on LPARs

When you assign adapters and domains to an LPAR on the HMC or SE, you indirectly assign virtual cryptographic devices.
For example, assigning adapter ID 00 and 02 as well as domains 0002, 0003, and 0005 to an LPAR implicitly assigns six virtual cryptographic devices to the LPAR: (00,0002), (00,0003), (00,0005), (02,0002), (02,0003), and (02,0005).

You can choose between two types of access to a cryptographic domain:

**To use cryptographic functions.**

A domain that is assigned to an LPAR for usage access is called a *usage domain* of that LPAR on the HMC or SE.

**To manage or control the domain, including the management of the master keys.**

A domain that is assigned to an LPAR for management (control) access is called a *control domain* of that LPAR on the HMC or SE.

Every usage domain of an LPAR must also be a control domain of that LPAR.

The list of usage domains and the list of adapter IDs define the list of virtual cryptographic devices that are assigned to an LPAR. For example, if 00 is an adapter ID and 0002 is a usage domain ID, then the virtual cryptographic device (00,0002) is assigned to the LPAR.

### Cryptographic devices on z/VM

In z/VM, the virtual cryptographic devices available to a guest are defined by using the `CRYPTO` directory statement:

- The `CRYPTO APDEDICATE` statement assigns domain IDs and adapter IDs to the guest. This assignment implicitly defines a list of dedicated virtual cryptographic devices. All virtual cryptographic devices that are determined by an ID from the adapter list of that guest and an ID from the domain list of that guest are assigned to the guest.

- The `CRYPTO APVIRT` statement assigns one virtual cryptographic device that can be shared among multiple guests with a guest-specific virtualized adapter ID and a virtualized domain ID.

Virtual cryptographic devices in z/VM can be either shared or dedicated, but not both.

Linux on z/VM with access to a shared cryptographic accelerator can either observe an accelerator or a CCA coprocessor.

When cryptographic coprocessors are shared, only clear-key RSA and random number functions are available to the Linux instance. Other requests are rejected by z/VM. For more information about supported functions, see the z/VM publications.

### Cryptographic devices on Linux

In Linux, virtual cryptographic devices are called *AP queues*. The name of an AP queue consists of two parts, the adapter ID and the domain ID, both in hexadecimal notation. For example, if cryptographic adapters with the IDs 00 and 02 are selected, and the domains IDs 0002, 0003 and 0005 have been configured on the cryptographic adapter, then the following AP queues are defined to Linux:
Setting up the cryptographic device driver

Configure the cryptographic device driver through the domain= and the poll_thread= module parameters. You might also have to set up libraries.

The cryptographic device driver consists of multiple, separate modules:

- **ap**  
  AP bus module.

- **zcrypt**  
  Cryptographic Coprocessor interface, Cryptographic Coprocessor message type 6, Cryptographic Coprocessor message type 50. Support for message type 6 includes secure key and RNG requests. Support for message type 50 includes RSA requests for both modulus-exponent and Chinese-Remainder Theorem variants.

- **zcrypt_cex4**  
  device driver for CEX6, CEX5, and CEX4 adapters. (CEX6 is recognized, but handled as a CEX5 adapter.)

- **zcrypt_cex2a**  
  device driver for CEX3A adapters.

- **zcrypt_pcixcc**  
  device driver for CEX3C adapters.

For information about setting up cryptographic hardware on your mainframe, see *zSeries Crypto Guide Update, SG24-6870*.

Module parameters

The cryptographic device driver consists of multiple, separate modules. You configure the device driver through module parameters when you load the AP bus module.

Use the `modprobe` command to load the AP bus module:

```
modprobe ap domain=autoselect
```

or

```
modprobe ap domain=<domain> poll_thread=0
```

or

```
modprobe ap domain=<domain> poll_thread=1
```

where
<domain>

is an integer that identifies the default cryptographic domain for the Linux instance. You define cryptographic domains in the LPAR activation profile on the HMC or SE.

The default value (domain=autoselect) causes the device driver to choose one of the available domains automatically.

**Important:** Be sure to enter an existing domain. The Trusted Key Entry (TKE) workstation does not find the cryptographic adapters if a non-existing domain is entered here. All CCA applications use the default domain, and do not work correctly if the specified domain does not exist.

<poll_thread>

is an integer argument and enables a polling thread to tune cryptographic performance. Valid values are 1 (enabled) or 0 (disabled, this value is the default). For details, see “Setting the polling thread” on page 434.

**Note:** If you are running Linux in an LPAR, AP interrupts are used instead of the polling thread. The polling thread is disabled when AP interrupts are available. See “Using AP adapter interrupts” on page 435.

All other modules are loaded automatically when they are required.

To remove a single module, for example, a module that supports a card type that is no longer available, issue a command of the following form:

```
# rmmod <module_name>
```

**Examples**

- This example loads the cryptographic device driver module ap if Linux runs in an LPAR with only one cryptographic domain:

```
# modprobe ap
```

- This example loads the cryptographic device driver module ap to operate within the cryptographic domain 1:

```
# modprobe ap domain=1
```

See the modprobe man page for command details.

**Accessing cryptographic devices**

User space programs access cryptographic devices through a single device node.

In Red Hat Enterprise Linux 7.5, udev creates the device node /dev/z90crypt for you. The device node is assigned to the miscellaneous devices.

**Accessing long random numbers**

Applications can access large amounts of random number data through a character device.

**Prerequisites:**
At least one cryptographic feature must be installed in the system and one coprocessor CEX6C, CEX5C, CEX4C, or CEX3C, must be configured. Alternatively, a z14 with TRNG support can be used as a back-end for the random-number character device.

Linux on z/VM needs a dedicated cryptographic coprocessor or a shared cryptographic device that is backed only by coprocessors. Automatic creation of the random number character device requires udev. The cryptographic device driver zcrypt must be loaded.

If the cryptographic device driver detects at least one coprocessor capable of generating long random numbers, a new miscellaneous character device is registered. The new device can be found under /proc/misc as hw_random. udev creates a character device called /dev/hwrng and a symbolic link called /dev/hw_random and pointing to /dev/hwrng.

Reading from the character device or the symbolic link returns the hardware-generated long random numbers. However, do not read excess amounts of random number data from this character device as the data rate is limited due to the cryptographic hardware architecture.

Removing the last available coprocessor adapter while the cryptographic device driver is loaded automatically removes the random number character device. Reading from the random number character device while all coprocessor adapters are set offline results in an input/output error (EIO). After at least one adapter is set online again, reading from the random number character device continues to return random number data.

### Working with cryptographic devices

Typically, cryptographic devices are not directly accessed by users but through user programs. Some tasks can be performed through the sysfs interface.

- Displaying information about cryptographic devices
- Setting devices online or offline on page 433
- Setting the polling thread on page 434
- Using AP adapter interrupts on page 435
- Setting the polling interval on page 435
- Dynamically adding and removing cryptographic adapters on page 436
- Displaying information about the AP bus on page 437
- Unloading the cryptographic device driver on page 438

### Displaying information about cryptographic devices

Use the `lszcrypt` command to display status information about your cryptographic devices; alternatively, you can use sysfs.

**About this task**

For information about `lszcrypt`, see Display cryptographic devices on page 591.

Each cryptographic adapter is represented in a sysfs directory of the form /sys/bus/ap/devices/card<XX>
where <XX> is the device index for each device. The valid device index range is hex 00 to hex 3f. For example, device 0x1a can be found under /sys/bus/ap/devices/card1a. The sysfs directory contains a number of attributes with information about the cryptographic adapter.

**Table 54. Cryptographic adapter attributes**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ap_functions</td>
<td>Read-only attribute that represents the function facilities that are installed on this device.</td>
</tr>
<tr>
<td>depth</td>
<td>Read-only attribute that represents the input queue length for this device.</td>
</tr>
<tr>
<td>hwtype</td>
<td>Read-only attribute that represents the hardware type for this device. The following values are defined:</td>
</tr>
<tr>
<td></td>
<td>8 CEX3A adapters.</td>
</tr>
<tr>
<td></td>
<td>9 CEX3C adapters.</td>
</tr>
<tr>
<td></td>
<td>10 CEX4A, CEX4C, or CEX4P adapters.</td>
</tr>
<tr>
<td></td>
<td>11 CEX5A, CEX5C, or CEX5P adapters.</td>
</tr>
<tr>
<td>raw_hwtype</td>
<td>Read-only attribute that represents the original hardware type of the cryptographic adapter.</td>
</tr>
<tr>
<td>modalias</td>
<td>Read-only attribute that represents an internally used device bus-ID.</td>
</tr>
<tr>
<td>online</td>
<td>Read-write attribute that shows whether the device is online (1) or offline (0).</td>
</tr>
<tr>
<td>pendingq_count</td>
<td>Read-only attribute that represents the number of requests in the hardware queue.</td>
</tr>
<tr>
<td>request_count</td>
<td>Read-only attribute that represents the number of requests that are already processed by this device.</td>
</tr>
<tr>
<td>requestq_count</td>
<td>Read-only attribute that represents the number of outstanding requests (not including the requests in the hardware queue).</td>
</tr>
<tr>
<td>type</td>
<td>Read-only attribute that represents the type of this device. The following types are defined:</td>
</tr>
<tr>
<td></td>
<td>• CEX3A, CEX3C</td>
</tr>
<tr>
<td></td>
<td>• CEX4A, CEX4C, CEX4P</td>
</tr>
<tr>
<td></td>
<td>• CEX5A, CEX5C, CEX5P</td>
</tr>
</tbody>
</table>

**Setting devices online or offline**

Use the `chzcrypt` command to set cryptographic devices online or offline.

**Procedure**

- Preferably, use the `chzcrypt` command with the `-e` option to set cryptographic devices online, or use the `-d` option to set devices offline.

**Examples:**

- To set cryptographic devices (in decimal notation) 0, 1, 4, 5, and 12 online issue:

  ```
  # chzcrypt -e 0 1 4 5 12
  ```

- To set all available cryptographic devices offline issue:
For more information about chzcrypt, see “chzcrypt - Modify the cryptographic configuration” on page 503.

- Alternatively, write 1 to the online sysfs attribute of a cryptographic device to set the device online, or write 0 to set the device offline.

Examples:
- To set a cryptographic device with device ID 0x3e online issue:
  ```bash
  # echo 1 > /sys/bus/ap/devices/card3e/online
  ```
- To set a cryptographic device with device ID 0x3e offline issue:
  ```bash
  # echo 0 > /sys/bus/ap/devices/card3e/online
  ```
- To check the online status of the cryptographic device with device ID 0x3e issue:
  ```bash
  # cat /sys/bus/ap/devices/card3e/online
  ```

The value is 1 if the device is online or 0 otherwise.

### Setting the polling thread

For Linux on z/VM, enabling the polling thread can improve cryptographic performance.

**About this task**

Linux in LPAR mode supports interrupts that indicate the completion of cryptographic requests. See “Using AP adapter interrupts” on page 435. If AP interrupts are available, it is not possible to activate the polling thread.

Depending on the workload, enabling the polling thread can increase cryptographic performance. For Linux on z/VM, the polling thread is deactivated by default.

The cryptographic device driver can run with or without the polling thread. When it runs with the polling thread, one processor constantly polls the cryptographic cards for finished cryptographic requests while requests are being processed. The polling thread sleeps when no cryptographic requests are being processed. This mode uses the cryptographic cards as much as possible, at the cost of blocking one processor during cryptographic operations.

Without the polling thread, the cryptographic cards are polled at a much lower rate. The lower rate results in higher latency, and reduced throughput for cryptographic requests, but without a noticeable processor load.

**Procedure**

- Use the chzcrypt command to set the polling thread.

**Examples:**
- To activate the polling thread issue:
To deactivate the polling thread issue:

```
# chzcrypt -n
```

For more information about `chzcrypt`, see “chzcrypt - Modify the cryptographic configuration” on page 503.

- Alternatively, you can set the polling thread through the `poll_thread` sysfs attribute. This read-write attribute can be found at the AP bus level.

**Examples:**

- To activate a polling thread for a device 0x3e issue:
  
  ```
  echo 1 > /sys/bus/ap/devices/card3e/poll_thread
  ```

- To deactivate a polling thread for a cryptographic device with bus device-ID 0x3e issue:
  
  ```
  echo 0 > /sys/bus/ap/devices/card3e/poll_thread
  ```

### Using AP adapter interrupts

To improve cryptographic performance for Linux instances that run in LPAR mode, use AP interrupts.

**About this task**

Using AP interrupts instead of the polling thread frees one processor while cryptographic requests are processed.

During initialization, the zcrypt device driver checks whether AP adapter interrupts are supported by the hardware. If so, polling is disabled and the interrupt mechanism is automatically used.

To query whether AP adapter interrupts are used, read the sysfs attribute `interrupt` of the device. Another interrupt attribute at the AP bus level, `/sys/bus/ap/ap_interrupts`, indicates that the AP bus is able to handle interrupts.

**Example**

To read the interrupt attribute for a device 0x3e with domain 0xf, issue:

```
# cat /sys/bus/ap/devices/card3e/3e.000f/interrupt
```

If interrupts are used, the attribute shows “Interrupts enabled”, otherwise “Interrupts disabled”.

### Setting the polling interval

Request polling is supported at nanosecond intervals.
Procedure

- Use the `lszcrypt` and `chzcrypt` commands to read and set the polling time.

Examples:
- To find out the current polling time, issue:
  ```
  # lszcrypt -b
  ...
  poll_timeout=250000 (nanoseconds)
  ```
- To set the polling time to 1 microsecond, issue:
  ```
  # chzcrypt -t 1000
  ```

For more information about `lszcrypt` and `chzcrypt`, see “`lszcrypt - Display cryptographic devices`” on page 591 and “`chzcrypt - Modify the cryptographic configuration`” on page 503.

- Alternatively, you can set the polling time through the `poll_timeout` sysfs attribute. This read-write attribute can be found at the AP bus level.

Examples:
- To read the `poll_timeout` attribute for the ap bus issue:
  ```
  # cat /sys/bus/ap/poll_timeout
  ```
- To set the `poll_timeout` attribute for the ap bus to poll, for example, every microsecond, issue:
  ```
  # echo 1000 > /sys/bus/ap/poll_timeout
  ```

Dynamically adding and removing cryptographic adapters

On an LPAR, you can add or remove cryptographic adapters without the need to reactivate the LPAR after a configuration change.

About this task

z/VM does not support dynamically adding or removing cryptographic adapters.

Linux attempts to detect new cryptographic adapters and set them online every time a configuration timer expires. Read or modify the expiration time with the `lszcrypt` and `chzcrypt` commands.

Adding or removing of cryptographic adapters to or from an LPAR is transparent to applications that use clear key functions. If a cryptographic adapter is removed while cryptographic requests are being processed, the device driver automatically resubmits lost requests to the remaining adapters. Special handling is required for secure key.

Secure key requests are submitted to a dedicated cryptographic coprocessor. If this coprocessor is removed or lost, new requests cannot be submitted to a different coprocessor. Therefore, dynamically adding and removing adapters with a secure
key application requires support within the application. For more information about secure key cryptography, see Secure Key Solution with the Common Cryptographic Architecture Application Programmer’s Guide, SC33-8294. You can obtain this book at

www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_cca.html

Alternatively, you can read or set the configuration timer through the config_time sysfs attribute. This read-write attribute can be found at the AP bus level. Valid values for the config_time sysfs attribute are in the range 5 - 120 seconds.

For the secure key cryptographic functions on CEX4P and CEX5P features, see Exploiting Enterprise PKCS #11 using openCryptoki, SC34-2713. You can obtain it at


**Procedure**

You can work with cryptographic adapters in the following ways:

- Add or remove cryptographic adapters by using the SE or HMC. After the configuration timer expires, the cryptographic adapter is added to or removed from Linux, and the corresponding sysfs entries are created or deleted.
- Enable or disable a cryptographic adapter by using the chzcrypt command. The cryptographic adapter is only set online or offline in sysfs. The sysfs entries for the cryptographic adapter are retained. Use the lszcrypt command to check the results of the chzcrypt command.

**Examples**

- To use the lszcrypt and chzcrypt commands to find out the current configuration timer setting, issue:

  ```
  # lszcrypt -b
  ...
  config_time=30 (seconds)
  ...
  
  ```

  In the example, the timer is set to 30 seconds.

- To set the configuration timer to 60 seconds, issue:

  ```
  # chzcrypt -c 60
  
  ```

To use sysfs to find out the current configuration timer setting, issue:

- To read the configuration timer setting, issue:

  ```
  # cat /sys/bus/ap/config_time
  
  ```

- To set the configuration timer to 60 seconds, issue:

  ```
  # echo 60 > /sys/bus/ap/config_time
  
  ```

**Displaying information about the AP bus**

Use the lszcrypt -b command to display status information about the AP bus; alternatively, you can use sysfs.
About this task

For information about `lszcrypt -b`, see “lszcrypt - Display cryptographic devices” on page 591.

The AP bus is represented in sysfs as a directory of the form
```
/sys/bus/ap
```

The sysfs directory contains a number of attributes with information about the AP bus.

Table 55. AP bus attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ap_domain</td>
<td>Read-write attribute that represents the default domain selected by the kernel. Alternatively, you can select the domain using a kernel parameter, or a module parameter during module load. See “Module parameters” on page 430.</td>
</tr>
<tr>
<td>ap_max_domain_id</td>
<td>Read-only attribute that represents the largest possible domain ID. Domain IDs can range from 0 to this number, which depends on the mainframe model.</td>
</tr>
<tr>
<td>ap_control_domain_mask</td>
<td>Read-only attribute that represents the installed control domain facilities as a 32-byte field in hexadecimal notation. A maximum number of 256 domains can be addressed. Each bit position represents a dedicated control domain.</td>
</tr>
<tr>
<td>ap_usage_domain_mask</td>
<td>Read-only attribute that represents the installed usage domain facilities as a 32-byte field in hexadecimal notation. A maximum number of 256 domains can be addressed. Each bit position represents a usage domain.</td>
</tr>
<tr>
<td>ap_interrupts</td>
<td>Read-only attribute that indicates whether interrupt handling for the AP bus is enabled.</td>
</tr>
<tr>
<td>config_time</td>
<td>Read-write attribute that represents a time interval in seconds used to detect new crypto devices.</td>
</tr>
<tr>
<td>poll_thread</td>
<td>Read-write attribute that indicates whether polling for the AP bus is enabled.</td>
</tr>
<tr>
<td>poll_timeout</td>
<td>Read-write attribute that represents the time interval of the poll thread in nanoseconds.</td>
</tr>
</tbody>
</table>

Example

```
# lszcrypt -b
ap_domain=0x6
ap_max_domain_id=0x54
ap_interrupts are enabled
config_time=30 (seconds)
poll_thread is disabled
poll_timeout=250000 (nanoseconds)
```

Unloading the cryptographic device driver

You can use `rmmod` to unload the cryptographic device driver modules.

Before you begin

The use count of the modules must be zero before you can unload them.
Procedure

- To unload the entire cryptographic device driver, explicitly unload each module. For example:

```
# rmmod zcrypt_pcixcc zcrypt_cex2a zcrypt_cex4 ap
```

- Alternatively, unload all unused modules that are related to zcrypt_api. You must unload only modules that were actually loaded. For example, if only the zcrypt_cex2a and zcrypt_cex4 modules are loaded in addition to ap use:

```
# rmmod zcrypt_cex2a zcrypt_cex4 ap
```

List the arguments in the order given.

External programming interfaces

Applications can directly access the cryptographic device driver through an API.

Programmers: This information is intended for those who want to program against the cryptographic device driver or against the available cryptographic libraries.

For information about the library APIs, see the following files in the Linux source tree:

- The libica library /usr/include/ica_api.h
- The openCryptoki library /usr/include/opencryptoki/pkcs11.h
- The CCA library /opt/IBM/CCA/include/csulincl.h
- The EP11 library /usr/include/ep11-host-devel/ep11.h and ep11adm.h

ep11.h and ep11adm.h are included in the EP11 devel package. pkcs11.h is included in the opencryptoki devel package. ica_api.h is provided by the libica-devel package. csulincl.h is present after the csulcca package is installed. Install the libica-devel-<version> RPM. You can obtain the openCryptoki package from sourceforge at: [sourceforge.net/projects/opencryptoki]

Clear key cryptographic functions

The libica library provides a C API to clear-key cryptographic functions that are supported by IBM Z hardware. You can configure both openCryptoki (by using the icatoken) and openssl (by using the ibmca engine) to use IBM Z clear-key cryptographic hardware support through libica. See *libica Programmer’s Reference, SC34-2602* for details about the libica functions.

If you must circumvent libica and access the cryptographic device driver directly, your user space program must open the zcrypt device node, and submit the cryptographic request with an IOCTL. The IOCTL subfunction ICARSAMODEXPO performs RSA modular exponent encryption and decryption. The IOCTL ICARSACRT performs RSA CRT decryption. See the cryptographic device driver header file in the Linux source tree: /usr/include/asm/zcrypt.h

Secure key cryptographic functions

To use secure key cryptographic functions in your user space program, see *Secure Key Solution with the Common Cryptographic Architecture Application Programmer’s*

To use secure key cryptographic functions in your user space program by accessing an EP11 coprocessor adapter, see *Exploiting Enterprise PKCS #11 using openCryptoki*, SC34-2713. You can obtain it at [www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/sec_hw_supp.html](http://www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/sec_hw_supp.html).

**Reading true random numbers**

To read true random numbers, a user space program must open the `hwrng` device and read as many bytes as needed from the device.

**Tip:** Using the output of the `hwrng` device to periodically reseed a pseudo-random number generator might be an efficient use of the random numbers.
Chapter 40. Pseudorandom number generator device driver

The pseudorandom number generator (PRNG) device driver provides user-space applications with pseudorandom numbers generated by the IBM Z CP Assist for Cryptographic Function (CPACF).

The PRNG device driver supports the Deterministic Random Bit Generator (DRBG) requirements that are defined in NIST Special Publication 800-90/90A. The device driver uses the SHA-512 based DRBG mechanism.

To use the SHA-512 algorithm, the device driver requires version 5 of the Message Security Assist (MSA), which is available as of the EC12 and zBC12 with the latest firmware level. During initialization of the prng kernel module, or, if prng is compiled into the kernel, during kernel startup, the device drivers checks for the prerequisite.

If the prerequisites for SHA-512 mode are not fulfilled, the device driver uses the Triple Data Encryption Standard (TDES) algorithm instead. In TDES mode, the PRNG device driver uses a DRBG in compliance with ANSI X9.17 based on the TDES cipher algorithm. You can force the fallback to TDES mode by using the prng.mode=1 kernel parameter or mode=1 module parameter.

Terminology hint: Various abbreviations are commonly used for Triple Data Encryption Standard, for example: TDES, triple DES, 3DES, and TDEA.

Setting up the pseudorandom number device driver

In Red Hat Enterprise Linux, the pseudorandom number device driver is compiled as a module. To use it, you must load the device driver module, and optionally make it available to non-root users.

Module parameters

You can load and configure the PRNG device driver if it was compiled as a separate module.

Module parameter syntax

```
modprobe prng -m=0 -c=256
```

where:

- `mode` specifies the operation mode (0 for SHA-512, 1 for TDES).
- `chunksize` specifies the chunk size.
- `reseed_limit` specifies the reseed limit.
mode=

specifies the mode in which the device driver runs:

0  Default. In this mode, the device driver automatically detects the MSA extension level and feature enablement. The device driver runs in SHA512 mode if the requirements are fulfilled, otherwise it falls back to TDES mode.

1  forces the device driver to run in TDES mode. The device driver starts only if the requirements for TDES mode are fulfilled.

2  forces the device driver to run in SHA512 mode. The device driver starts only if the requirements for SHA512 mode are fulfilled. The device driver does not fall back to TDES mode.

<sizeparam>

adjusts the random-buffer block size that the device driver uses to generate new random bytes. In TDES mode, this value can be in the range 8 - 65536, for SHA512 mode, the range is 64 - 65536. The default is 256 bytes.

<reseedparam>

adjusts the reseed limit in SHA512 mode. Multiply this value with the chunksize to obtain the reseed boundary in bytes. The value can be in the range 10000 - 100000. The default is 100000. In TDES mode, the reseed limit is a constant value of 4096 bytes.

Controlling access to the device node

Red Hat Enterprise Linux by default assigns access mode 0644 to /dev/prandom.

To restrict access to the device node to root users, add the following udev rule. It prevents non-root users from reading random numbers from /dev/prandom.

```
KERNEL=="prandom", MODE="0600", OPTIONS="last_rule"
```

If access to the device is restricted to root, add the following udev rule. It automatically extends access to the device to other users.

```
KERNEL=="prandom", MODE="0644", OPTIONS="last_rule"
```

Working with the PRNG device driver

Read random numbers and control the settings of the PRNG device driver.

Tasks include:

- “Reading pseudo-random numbers”
- “Displaying PRNG information” on page 443
- “Reseeding the PRNG” on page 444
- “Setting the reseed limit” on page 444

Reading pseudo-random numbers

The pseudo-random number device is read-only. Use the read function, cat program, or dd program to obtain random numbers.

Example

In this example bs specifies the block size in bytes for transfer, and count specifies the number of records with block size. The bytes are written to the output file.
Displaying PRNG information
Read the attributes of the prandom device in sysfs.

About this task
The sysfs representation of a PRNG device is a directory: /sys/devices/virtual/misc/prandom. This sysfs directory contains a number of attributes with information about the device.

Table 56. Attributes with PRNG information

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>chunksize</td>
<td>The size, in bytes, of the random-data bytes buffer that is used to generate new random numbers. The value can be in the range 64 bytes - 64 KB. The default is 256 bytes. It is rounded up to the next 64-byte boundary and can be adjusted as a module parameter when you start the module.</td>
</tr>
<tr>
<td>byte_counter</td>
<td>The number of random bytes generated since the PRNG device driver was started. You can reset this value only by removing and reloading the kernel module, or rebooting Linux (if PRNG was compiled into the kernel). This attribute is read-only.</td>
</tr>
</tbody>
</table>
| errorflag       | SHA512 mode only: 0 if the PRNG device driver is instantiated and running well. Any other value indicates a problem. If there is an error indication other than 0:  
                   • The DRBG does not provide random data bytes to user space  
                   • The read() function fails  
                   • The error code errno is set to EPIPE (broken pipe)  
                   This attribute is read-only. |
| mode            | SHA512 if the PRNG device driver runs in SHA512 mode, TDES if the PRNG device driver runs in TDES mode. This attribute is read-only.            |
| reseed          | SHA512 mode only: An integer, writable only by root. Write any integer to this attribute to trigger an immediate reseed of the PRNG. See “Reseeding the PRNG” on page 444. |
| reseed_limit    | SHA512 mode only: An integer, writable only by root to query or set the reseed counter limit. Valid values are in the range 10000 - 100000. The default is 100000. See “Setting the reseed limit” on page 444. |
| strength        | SHA512 mode only: A read-only integer that shows the security strength according to NIST SP800-57. Returns the integer value of 256 in SHA512 mode. |

Procedure

Issue a command of this form to read an attribute:

```
# cat /sys/devices/virtual/misc/prandom/<attribute>
```

where `<attribute>` is one of the attributes of Table 56.

Example

This example shows a prandom device that is running in SHA512 mode, set to reseed after 2.56 MB:
Setting the reseed limit

The PRNG reseeds after chunksize \( \times \text{reseed\_limit} \) bytes are read. By default, 100000 \( \times 256 = 25.6 \) MB can be read before an automatic reseed occurs.

Procedure

To set the number of times a chunksize amount of random data can be read from the PRNG before reseeding, write the number to the reseed\_limit attribute. For example:

```
# echo 10000 > /sys/devices/virtual/misc/prandom/reseed\_limit
```

The reseed\_limit value must be in the range 10000 - 100000.

Reseeding the PRNG

You can force a reseed by writing to the reseed attribute.

Procedure

To reseed the PRNG, write an integer to its reseed attribute:

```
# echo 1 > /sys/devices/virtual/misc/prandom/reseed
```

Writing any integer value to this attribute triggers an immediate reseed of the PRNG instance.
Chapter 41. True random-number generator device driver

The true random number generator (TRNG) device driver provides user-space applications with random data generated from the IBM Z hardware CPACF true random source.

Setting up the TRNG device driver

The true random-number generator can be built into the kernel or compiled as a separate module.

The true random-number generator requires Message-Security-Assist Extension 7 (MSA 7), which is available as of the IBM z14. During initialization of the TRNG kernel module, or, if TRNG is compiled into the kernel, during kernel startup, the device drivers checks for the prerequisite. If the prerequisite is not fulfilled, the device driver silently exits.

If the TRNG device driver is compiled as a module, the module registers itself to the CPU feature MSA. You can activate the device driver manually with the command modprobe s390-trng. If the TRNG device driver is built into the kernel, no activation is necessary.

The kernel module trng-s390 is located with the character device drivers under /drivers/char/hw_random.

There are no kernel or module parameters for the TRNG device driver.

Device nodes for random data

The true random-number generator device driver provides two interfaces to user space applications: the device node /dev/trng for direct access, and the generic device node /dev/hwrng if available. The /dev/hwrng node appears if at least one random data source successfully registered with the hwrng device driver. The TRNG registers as such a random datasource.

As of the z14, the kernel random device also uses the CPACF TRNG true random source through the arch_get_random_* functions. The kernel random device provides two device nodes, /dev/random and /dev/urandom. The arch_get_random_* functions require the CPACF TRNG. If the prerequisite is not fulfilled, the functions return false to indicate that no random data could be produced.

Working with the TRNG device driver

Read random numbers and retrieve the counters of the TRNG device driver.

Tasks include:

- “Reading random numbers” on page 446
- “Displaying TRNG information” on page 446
Reading random numbers

The TRNG device is read-only. Use the read function, cat program, or dd program to obtain random numbers.

Example

In this example bs specifies the block size in bytes for transfer, and count specifies the number of records with block size. The bytes are written to the output file.

```
dd if=/dev/trng of=<output file name> bs=<xxxx> count=<nnnn>
```

Displaying TRNG information

Read the byte_counter attribute of the TRNG device in sysfs.

About this task

The sysfs representation of a TRNG device is a directory: /sys/devices/virtual/misc/trng. This sysfs directory contains an attribute, byte_counter, with statistical data.

Procedure

Issue this command to read the byte_counter attribute:

```
# cat /sys/devices/virtual/misc/trng/byte_counter
```

Example

To see statistics of a TRNG device, issue:

```
# cat /sys/devices/virtual/misc/trng/byte_counter
trng: 6187
hwrng: 528
arch: 1319696
total: 1326411
```

Where:

- **trng** shows the number of bytes delivered through the /dev/trng device node.
- **hwrng** shows the bytes retrieved from the generic hw_rng device driver and contributed to /dev/hwrng.
- **arch** shows the amount of data that is received through the arch random implementation and delivered to the random device driver device nodes /dev/random and /dev/urandom.
- **total** shows the sum of all bytes.
Chapter 42. Protected key device driver

The protected key device driver provides functions for generating and verifying protected keys.

Protected keys are encrypted with wrapping keys that, for Linux in LPAR mode, are specific to the LPAR. For guests of z/VM or KVM, the wrapping key is specific to the guest. Both the wrapping keys and the clear key values of protected keys are invisible to the operating system. Protected keys are designed for accelerated encryption and decryption with CPACF. For more information, see the chapter about protected keys in z/Architecture Principles of Operation, SA22-7832.

The protected key device driver requires the message-security-assist-extension 3 facility (MSA level 3), which was introduced with z196.

The device driver provides the following functions to cryptographic applications:

- Generate a random secure key, then generate a protected key from the secure key.
  The secure key must be available to create a new version of the protected key whenever the current protected key is invalidated.

- Generate a secure key from a clear key, then generate a protected key from the secure key.
  The clear key must be in memory when the protected key is generated.
  Thereafter, the clear key can be deleted.
  The secure key must be available to create a new version of the protected key whenever the current protected key is invalidated.

- Generate a protected key from a clear key. The clear key must be in memory when the protected key is generated.
  The clear key must also be available to create a new protected key if the existing protected key is invalidated.

Secure keys are encoded with a master key that is held in Crypto Express® adapters. Functions that involve secure keys require an IBM Crypto Express adapter in CCA coprocessor mode with a valid master key. For Linux on z/VM, the adapter must be dedicated to the z/VM guest virtual machine.

Loading the device driver module

The protected key device driver is compiled as a separate module, pkey. This module is required for working with protected keys.

Load the pkey module with the `modprobe` command. The pkey module has no module parameters.

```bash
# modprobe pkey
```

External programming interfaces

Applications can use the protected key device driver through ioctl's or corresponding kernel APIs.
Programmers: This information is intended for programmers of cryptographic applications who want to use protected keys for accelerated cryptographic operations with CPACF.

Issue ioctls on the misc character device /dev/pkey to generate and handle protected keys. The ioctl interface, including the required defines and structure definitions, is described in arch/s390/include/uapi/asm/pkey.h. Each ioctl has a matching kernel API that is also described in this file.

<p>| Table 57. ioctls of the protected key device driver |</p>
<table>
<thead>
<tr>
<th>Name</th>
<th>Structure passed</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKEY_GENSECK</td>
<td>struct pkey_genseck</td>
<td>Obtain a random secure key from a cryptographic coprocessor. The secure key is encrypted with the master key of the coprocessor.</td>
</tr>
<tr>
<td>PKEY_CLR2SECK</td>
<td>struct pkey_clr2seck</td>
<td>Obtain a secure key from a cryptographic coprocessor. The secure key is generated from an existing clear key and encrypted with the master key of the coprocessor.</td>
</tr>
<tr>
<td>PKEY_SEC2PROTK</td>
<td>struct pkey_sec2protk</td>
<td>Obtain a protected key from a cryptographic coprocessor. The protected key is generated from an existing secure key.</td>
</tr>
<tr>
<td>PKEY_CLR2PROTK</td>
<td>struct pkey_clr2protk</td>
<td>Obtain a protected key from a cryptographic coprocessor. The protected key is generated from an existing clear key.</td>
</tr>
<tr>
<td>PKEY_FINDCARD</td>
<td>struct pkey_findcard</td>
<td>Find a cryptographic coprocessor that holds the applicable master key for an existing secure key.</td>
</tr>
<tr>
<td>PKEY_SKEY2PKEY</td>
<td>struct pkey_skey2pkey</td>
<td>Find a cryptographic coprocessor that holds the applicable master key for an existing secure key. Then use the coprocessor to obtain a protected key that is generated from the secure key. This ioctl call combines PKEY_FINDCARD and PKEY_SEC2PROTK.</td>
</tr>
</tbody>
</table>
Chapter 43. Hardware-accelerated in-kernel cryptography

The Linux kernel implements cryptographic operations for kernel subsystems like dm-crypt and IPSec. Applications can use these operations through the kernel cryptographic API.

In-kernel cryptographic operations can be performed by platform-specific implementations instead of the generic implementations within the Linux kernel.

On IBM Z, hardware-accelerated processing is available for some of these operations.

Hardware dependencies and restrictions

The cryptographic operations that can be accelerated by hardware implementations depend on your IBM Z hardware features and mode of operating Red Hat Enterprise Linux 7.5.

z196 and later IBM Z hardware supports hardware-acceleration for operations that cover the following standards:

- SHA-1
- SHA-256
- SHA-512
- DES and TDES (ECB, CBC, and CTR modes)
- AES (ECB, CBC, and CTR modes for all AES key sizes; XTS for 256-bit and 512-bit keys)
- GHASH
- PAES (AES with protected key)

CPACF dependencies

Hardware-acceleration for DES, TDES, AES, GHASH, and PAES requires the Central Processor Assist for Cryptographic Function (CPACF). Read the features line from /proc/cpuinfo to find out whether the CPACF feature is enabled on your hardware.

Example:

```
# cat /proc/cpuinfo | grep features
features : esan3 zarch stfle msa ldisp eimm dfp edat etf3eh highgprs te vx sie
```

In the output line, msa indicates that the CPACF feature is enabled. For information about enabling CPACF, see the documentation for your IBM Z hardware.

FIPS restrictions of the hardware capabilities

If the kernel runs in Federal Information Processing Standard (FIPS) mode, only FIPS 140-2 approved algorithms are available. DES, for example, is not approved by FIPS 140-2.
Read `/proc/sys/crypto/fips_enabled` to find out whether your kernel runs in FIPS mode.

**Example:**

```
# cat /proc/sys/crypto/fips_enabled
0
```

The kernel of the example does not run in FIPS mode. For kernels that run in FIPS mode, the output of the command is 1.

You control the FIPS mode with the `fips` kernel parameter, see "fips - Run Linux in FIPS mode" on page 657.

For more information about FIPS, see [src.nist.gov/publications/PubsFIPS.html](https://src.nist.gov/publications/PubsFIPS.html).

---

### Loading the support modules

Load support modules to enable hardware-acceleration for specific cryptographic operations. None of these modules have module parameters.

- **sha512_s390**
  enables hardware-acceleration for SHA-384 and SHA-512 operations.
  sha512_s390 requires the sha_common module.

- **ghash_s390**
  enables hardware-acceleration for Galois hashes.

- **aes_s390**
  enables hardware-acceleration for AES encryption and decryption for the following modes of operation:
  - ECB, CBC, and CTR for key lengths 128, 192, and 256 bits
  - XTS for key lengths 128 and 256 bits

- **des_s390**
  enables hardware-acceleration for DES and TDES for the following modes of operation: ECB, CBC, and CTR.

- **paes_s390**
  enables protected key AES encryption and decryption for the following modes of operation:
  - ECB, CBC, and CTR for key lengths 128, 192, and 256 bits
  - XTS for key lengths 128 and 256 bits

  The paes_s390 module requires the pkey device driver, see Chapter 42, "Protected key device driver," on page 447. The module also requires a cryptographic adapter in coprocessor mode for the creation and handling of protected keys.

Modules for SHA-1 and SHA-256 operations are built into the kernel.

**Note:** To use CPACF for AES-GCM operations, you must load both the aes_s390 and ghash_s390 module.

**Tip:** Load the modules with `modprobe`. `modprobe` handles dependencies on other modules for you.

**Example:**
Confirming hardware support for cryptographic operations

Read /proc/crypto to confirm that cryptographic operations are performed with hardware support.

Procedure

Read the driver lines from the content of /proc/crypto.

Example:

```
# cat /proc/crypto | grep driver
driver : sha512-s390
driver : sha224-s390
driver : sha256-s390
driver : sha1-s390
driver : ghash-s390
...
```

Each line that ends in -s390 indicates hardware-acceleration for a corresponding algorithm or mode.
Chapter 44. Instruction execution protection

The instruction execution protection feature on IBM mainframes protects against data execution, similar to the NX feature on other architectures.

Instruction execution protection prevents stack-overflow exploits and generally makes a system insensitive to buffer-overflow attacks.

Data instruction protection is available on IBM mainframe hardware with the IEP feature. For Linux as a guest of a hypervisor, the hypervisor must support and use the instruction execution protection feature. For example, Linux as a guest of KVM must run on a KVM host with kernel 4.11 or later, and with the IEP feature enabled.

Setting up instruction execution protection

By default, Linux uses the instruction execution protection feature if it is available. You can use the noexec kernel parameter to disable the feature in Linux.

Kernel parameter for instruction execution protection

| noexec=on  |
| noexec=off |

If set to on, noexec enables instruction execution protection, this is the default. If set to off, noexec disables instruction execution protection.

Controlling stack execution protection

To prevent stack-overflow exploits, the stack of a binary or shared library must be marked as not executable.

About this task

Use the execstack command to set, clear, or query the executable stack flag of ELF binaries and shared libraries (GNU_STACK). The execstack command is available from the prelink package. For details about execstack, see the man page.

Example

- Set and query the executable stack flag.

  ```
  # execstack -s /usr/bin/find
  # execstack -q /usr/bin/find
  X /usr/bin/find
  ```

  The leading X at the beginning of the query output line indicates that the stack is executable.
**Hint:** You can also use the `readelf` command to confirm that the stack can be executed.

```bash
# readelf -a /usr/bin/find | grep GNU_STACK -A 1
GNU_STACK 0x0000000000000000 0x0000000000000000 0x0000000000000000 0x0000000000000000 0x0000000000000000 0x0000000000000000 RWE 8
```

The **RWE** towards the end of the output line means read/write/execute. You can obtain the `readelf` command as part of the `binutils` package. For command details, see the man page.

- Clear and query the executable stack flag.

```bash
# execstack -c /usr/bin/find  
# execstack -q /usr/bin/find  
- /usr/bin/find
```

The leading `-` at the beginning of the query output line indicates that the stack is not executable.

**Hint:** You can also use the `readelf` command to confirm that the stack cannot be executed.

```bash
# readelf -a /usr/bin/find | grep GNU_STACK -A 1
GNU_STACK 0x0000000000000000 0x0000000000000000 0x0000000000000000 0x0000000000000000 0x0000000000000000 0x0000000000000000 RW 8
```

The **RW** towards the end of the output line means read/write, but not execute.
The Z hardware provides performance data that can be accessed by Linux. Gathering performance data constitutes an additional load on the Linux instance on which the application to be analyzed runs. Hardware support for data gathering can reduce the extra load and can yield more accurate data.

For the performance measurement facilities of z/VM, see “Performance monitoring for z/VM guest virtual machines” on page 371.

Other performance relevant information is provided in the context of the respective device driver or feature. For example, see “Working with DASD statistics in debugfs” on page 144 for DASD performance and “Starting and stopping collection of QETH performance statistics” on page 257 for qeth group devices.

Newest version

You can find the newest version of this book at www.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_distlibs.html

Restrictions

For prerequisites and restrictions see the IBM Z architecture specific information in the Red Hat Enterprise Linux 7.5 release notes at https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux
Chapter 45. Channel measurement facility

The IBM Z architecture provides a channel measurement facility to collect statistical data about I/O on the channel subsystem.

Data collection can be enabled for all CCW devices. User space applications can access this data through the sysfs.

The channel measurement facility provides the following features:

- Basic channel measurement format for concurrently collecting data on up to 4096 devices. (Specifying 4096 or more channels causes high memory consumption, and enabling data collection might not succeed.)
- Extended channel measurement format for concurrently collecting data on an unlimited number of devices.
- Data collection for all channel-attached devices, except those using QDIO (that is, except qeth and SCSI-over-Fibre channel attached devices)

Setting up the channel measurement facility

Configure the channel measurement facility by adding parameters to the kernel parameter file.

Channel measurement facility kernel parameters

```
| cmf.format=1024 |
| cmf.format=0    |
| cmf.format=-1   |
| cmf.maxchannels=1024 |
| cmf.maxchannels=<no_channels> |
```

Notes:
1. If you specify both parameter=value pairs, separate them with a blank.

where:

**cmf.format**

defines the format, 0 for basic and 1 for extended, of the channel measurement blocks. The default, -1, assigns a format depending on the hardware, the extended format for zEnterprise mainframes.

**cmf.maxchannels=<no_channels>**

limits the number of devices for which data measurement can be enabled concurrently with the basic format. The maximum for `<no_channels>` is 4096. A warning is printed if more than 4096 channels are specified. The channel measurement facility might still work; however, specifying more than 4096 channels causes a high memory consumption.

For the extended format, there is no limit and any value you specify is ignored.
Working with the channel measurement facility

Typical tasks that you need to perform when you work with the channel measurement facility is controlling data collection and reading data.

Enabling, resetting, and switching off data collection

Control data collection through the cmb_enable sysfs attribute of the device.

Procedure

Use a device’s cmb_enable attribute to enable, reset, or switch off data collection.

- To enable data collection, write 1 to the cmb_enable attribute. If data collection was already enabled, writing 1 to the attribute resets all collected data to zero.

   Issue a command of this form:

   ```
   # echo 1 > /sys/bus/ccw/devices/<device_bus_id>/cmb_enable
   ```

   where /sys/bus/ccw/devices/<device_bus_id> represents the device in sysfs.

   When data collection is enabled for a device, a subdirectory /sys/bus/ccw/devices/<device_bus_id>/cmf is created that contains several attributes. These attributes contain the collected data (see “Reading data”).

- To switch off data collection issue a command of this form:

   ```
   # echo 0 > /sys/bus/ccw/devices/<device_bus_id>/cmb_enable
   ```

   When data collection for a device is switched off, the subdirectory /sys/bus/ccw/devices/<device_bus_id>/cmf and its content are deleted.

Example

In this example, data collection for a device /sys/bus/ccw/devices/0.0.b100 is already active and reset:

```
# cat /sys/bus/ccw/devices/0.0.b100/cmb_enable
1
# echo 1 > /sys/bus/ccw/devices/0.0.b100/cmb_enable
```

Reading data

Read the sysfs attributes with collected I/O data, for example with the cat command.

Procedure

While data collection is enabled for a device, the directories that represent it in sysfs contain a subdirectory, cmf, with several read-only attributes. These attributes hold the collected data.

To read one of the attributes, issue a command of this form:

```
# cat /sys/bus/ccw/devices/<device_bus_id>/cmf/<attribute>
```

where /sys/bus/ccw/devices/<device_bus_id> is the directory that represents the device, and <attribute> the attribute to be read. Table 58 on page 459 summarizes the available attributes.
Table 58. Attributes with collected I/O data

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssch_rsch_count</td>
<td>An integer that represents the ssch rsch count value.</td>
</tr>
<tr>
<td>sample_count</td>
<td>An integer that represents the sample count value.</td>
</tr>
<tr>
<td>avg_device_connect_time</td>
<td>An integer that represents the average device connect time, in nanoseconds, per sample.</td>
</tr>
<tr>
<td>avg_function_pending_time</td>
<td>An integer that represents the average function pending time, in nanoseconds, per sample.</td>
</tr>
<tr>
<td>avg_device_disconnect_time</td>
<td>An integer that represents the average device disconnect time, in nanoseconds, per sample.</td>
</tr>
<tr>
<td>avg_control_unit_queuing_time</td>
<td>An integer that represents the average control unit queuing time, in nanoseconds, per sample.</td>
</tr>
<tr>
<td>avg_initial_command_response_time</td>
<td>An integer that represents the average initial command response time, in nanoseconds, per sample.</td>
</tr>
<tr>
<td>avg_device_active_only_time</td>
<td>An integer that represents the average device active only time, in nanoseconds, per sample.</td>
</tr>
<tr>
<td>avg_device_busy_time</td>
<td>An integer that represents the average device busy time, in nanoseconds, per sample.</td>
</tr>
<tr>
<td>avg_utilization</td>
<td>A percent value that represents the fraction of time that was spent in device connect time, plus function pending time, plus device disconnect time during the measurement period.</td>
</tr>
<tr>
<td>avg_sample_interval</td>
<td>An integer that represents the average time, in nanoseconds, between two samples during the measurement period. Can be “-1” if no measurement data was collected.</td>
</tr>
<tr>
<td>avg_initial_command_response_time</td>
<td>An integer that represents the average time in nanoseconds between the first command of a channel program being sent to the device and the command being accepted. Available in extended format only.</td>
</tr>
<tr>
<td>avg_device_busy_time</td>
<td>An integer that represents the average time in nanoseconds of the subchannel being in the “device busy” state when initiating a start or resume function. Available in extended format only.</td>
</tr>
</tbody>
</table>

**Example**

To read the avg_device_busy_time attribute for a device /sys/bus/ccw/devices/0.0.b100:

```
# cat /sys/bus/ccw/devices/0.0.b100/cmf/avg_device_busy_time
21
```
Chapter 46. OProfile hardware sampling support

OProfile is a performance analysis tool for Linux that can use hardware sampling support to capture performance data for processes, shared libraries, the kernel, and device drivers.

For general information about OProfile, see sourceforge.net/projects/oprofile

OProfile hardware sampling can be used for Linux instances in LPAR mode.

Note: OProfile and perf-based sampling tools use the CPU-measurement sampling facility and, therefore, cannot simultaneously collect sample data.

Setting up OProfile support

After you install the OProfile package that is provided with Red Hat Enterprise Linux, you must initialize OProfile on your Linux instance. Then, enable hardware sampling for the LPAR in which the Linux instance runs.

Initializing OProfile

Before initialization, the /dev/oprofile file system is not available and commands that act on files within this file system fail.

Issue:

```
# opcontrol --init
```

This command loads the oprofile module and initializes the OProfile support. For more information, see oprofile.sourceforge.net/docs

Setting up an LPAR for hardware sampling

To enable hardware sampling for an LPAR you must activate the LPAR with authorization for basic sampling control.

See the Support Element Operations Guide for your mainframe system for more information.

To check if hardware sampling is enabled, read the hwsampler attribute:

```
# cat /dev/oprofile/hwsampling/hwsampler
1
```

If hardware sampling is enabled, the value is 1.

If the value is 0, timer-interrupt based sampling is used. The reason might be that your IBM Z hardware does not support hardware sampling, that your LPAR has not been set up for hardware sampling, or that your Linux instance runs as a z/VM guest.

You can disable hardware sampling by writing 0 to the hwsampler attribute:
Working with OProfile

You might have to set the sampling interval and the sampler memory, and you might have to start and stop sampling.

- “Starting and stopping sampling”
- “Setting the sampling interval”
- “Setting the sampler memory”

Starting and stopping sampling

You start and stop sampling as you would on any hardware platform.

See [oprofile.sourceforge.net/docs](http://oprofile.sourceforge.net/docs) for details.

Setting the sampling interval

Set the sampling interval through the `/dev/oprofile/hwsampling/hw_interval` attribute in the `/dev/oprofile` file system.

Procedure

Issue a command of this form to set the sample interval:

```
# echo <value> > /dev/oprofile/hwsampling/hw_interval
```

where `<value>` is the sample interval in processor cycles. The sample interval must not exceed the value of the `hw_max_interval` attribute and it must not be smaller than the value of the `hw_min_interval` attribute. The default is 4096.

Example

This example sets the sampling rate to twice the default rate:

```
# echo 2048 > /dev/oprofile/hwsampling/hw_interval
```

Setting the sampler memory

Set the sampler memory size through the `/dev/oprofile/hwsampling/ hw_sdbt_blocks` attribute in the `/dev/oprofile` file system.

About this task

The best size for the sampler memory depends on the particular system and the workload to be measured. Providing the sampler with too little memory results in lost samples. Reserving too much system memory for the sampler impacts the overall performance and, hence, also the workload to be measured.

Procedure

To set the size of the memory that is reserved for sampled data, issue a command of this form:
where `<value>` is the memory size in multiples of 2 MB. The default is 1.

**Example**

```
# echo 2 > /dev/oprofile/hwsampling/hw_sdbt_blocks
```
Chapter 47. Using the CPU-measurement facilities

Use the CPU-measurement counter facility and sampling facility to obtain performance data for Linux in LPAR mode.

Counter facility
The hardware counters are grouped into the following counter sets:
• Basic counter set
• Problem-state counter set
• Crypto-activity counter set
• Extended counter set

A further common counter set, the Coprocessor group counter set, cannot be accessed from Linux on Z.

Sampling facility
The sampling facility includes the following sampling modes:
• Basic-sampling mode
• Diagnostic-sampling mode

The diagnostic-sampling mode is intended for use by IBM support only.

Conflict with OProfile: Perf-based sampling tools and OProfile use the CPU-measurement sampling facility and, therefore, cannot simultaneously collect sample data.

The number and type of individual counters and the details of the sampling facility depend on your Z hardware model. Use the lscpumf command to find out what is available for your hardware (see "lscpumf - Display information about the CPU-measurement facilities" on page 564). For details, see IBM The CPU-Measurement Facility Extended Counters Definition for z10™, z196, z114 and zEC12, SA23-2261.

You can use the perf tool on Linux to access the hardware counters and sample data of the CPU-measurement facilities. To use the perf tool, you need to install the perf tool package, perf-<version>.s390x.rpm, provided with Red Hat Enterprise Linux.

If you want to write your own application for analyzing counter or sample data, you can use the libpfm4 library. This library is available on sourceforge at perfmon2.sourceforge.net.

Working with the CPU-measurement counter facility
You can use the perf tool to work with the CPU-measurement counter facility for authorized LPARs.
• “Authorizing an LPAR for CPU-measurement counter sets” on page 466
• “Reading CPU-measurement counters for an application” on page 466
• “Collecting CPU-measurement sample data” on page 467
• “Setting limits for the sampling facility buffer” on page 468
• “Obtaining debug information” on page 469
Authorizing an LPAR for CPU-measurement counter sets

The LPAR within which the Linux instance runs must be authorized to use the CPU-measurement counter sets. Use the HMC or SE to authorize the LPAR for the counter sets you need.

Procedure

Perform these steps on the HMC or SE to grant authorization:
1. Navigate to the LPAR for which you want to grant authorization for the counter sets.
2. Within the LPAR profile, select the Security page.
3. Within the counter facility options, select each counter set you want to use. The coprocessor group counter set is not supported by Linux on Z.
4. Click Save.

What to do next

Deactivate, activate, and IPL the LPAR to make the authorization take effect. For more information, see the Support Element Operations Guide for your mainframe system.

Reading CPU-measurement counters for an application

Use the perf tool to read CPU-measurement counters with the scope of an application.

Before you begin

You must know the hexadecimal value of the counter number. You can find the decimal values in z/Architecture The Load-Program-Parameter and the CPU-Measurement Facilities, SA23-2260 and in IBM The CPU-Measurement Facility Extended Counters Definition for z10, z196, z114 and zEC12, SA23-2261.

Procedure

Issue a command of this form to read a counter:

```
# perf stat -e r<hex_counter_number> -- <path_to_app>
```

Where:

- `-e r<hex_counter_number>`
  specifies the hexadecimal value for the counter number as a raw event.

  **Tip**: You can read multiple counters by specifying a comma-separated list of raw events, for example, `-e r20,r21`.

- `<path_to_app>`
  specifies the path to the application to be evaluated. The counters are incremented for all threads that belong to the specified application.

For more information about the `perf` command, see the `perf` or `perf-stat` man page.
**Example**

To read the counters with hexadecimal values 20 (problem-state cycle count) and 21 (problem-state instruction count) for an application `/bin/df`:

```
# perf stat -e r20,r21 -- /bin/df
```

Filesystem 1K-blocks Used Available Use% Mounted on
/dev/dasda1 7188660 2521760 4306296 37% /
none 923428 88 923340 1% /dev/shm
/dev/dasdb1 7098728 2631972 4160152 40% /root

Performance counter stats for `/bin/df`:

```
1185753 raw 0x20
257509 raw 0x21
0.002507687 seconds time elapsed
```

**Collecting CPU-measurement sample data**

Use the `perf` tool to read CPU-measurement sample data.

**Procedure**

Issue a command of this form to read sample data:

```
# perf record -e cpum_sf/event=SF_CYCLES_BASIC/ -- <path_to_app>
```

Where `<path_to_app>` is the path to the application for which you want to collect sample data. If you specify `-a` instead of the double hyphen and path, system-wide sample data is collected. Instead of the symbolic name, you can also specify the raw event name `rB0000`.

**Example**

```
# perf record -e cpum_sf/event=SF_CYCLES_BASIC/ -- /bin/df
```

Filesystem 1K-blocks Used Available Use% Mounted on
/dev/dasda1 6967656 4132924 2474128 63% /root
[ perf record: Woken up 1 times to write data. ]
[ perf record: Captured and wrote 0.001 MB perf.data (~20 samples) ]

**What to do next**

You can now display the sample data by issuing the following command:

```
# perf report
```

For more information about collecting and displaying sample data with the `perf` command, see the `perf-record` and the `perf-report` man pages.

**Hint:** You can use the `perf record -F` option to collect sample data at a high frequency or the `perf record -c` option to collect sample data for corresponding short sampling intervals. Specified values must be supported by both the CPU-measurement sampling facility and `perf`. Issue `lscpu` `-f` to find out the maximum and minimum values for the CPU-measurement sampling facility. If `perf`
fails at a high sampling frequency, you might have to adjust the
kernel.perf_event_max_sample_rate system control to override default perf
limitations.

Setting limits for the sampling facility buffer
Use the `chcpumf` command to set the minimum and maximum buffer size for the
CPU-measurement sampling facility.

See “chcpumf - Set limits for the CPU measurement sampling facility buffer” on
page 496.

Before you begin

For each CPU, the CPU-measurement sampling facility has a buffer for writing
sample data. The required buffer size depends on the sampling function and the
sampling interval that is used by the perf tool. The sampling facility starts with an
initial buffer size that depends on the expected requirements, your IBM Z
hardware, and the available hardware resources. During the sampling process, the
sampling facility increases the buffer size if required.

The sampling facility is designed for autonomous buffer management, and you do
not usually need to intervene. You might want to change the minimum or
maximum buffer size, for example, for one of the following reasons:

- There are considerable resource constraints on your system that cause perf
  sampling to malfunction and sample data to be lost.
- As an expert user of perf and the sampling facility, you want to explore results
  with particular buffer settings.

Procedure

Use the `chcpumf` command to set the minimum and maximum buffer sizes.

1. Optional: Specify the `1scpumf` command with the `-i` parameter to display the
current limits for the buffer size (see “1scpumf - Display information about the
CPU-measurement facilities” on page 564).

2. Optional: Specify the `chcpumf` command with the `-m` parameter to set the
minimum buffer size.

Example:

```
# chcpumf -m 500
```

The value that you specify with `-m` is the minimum buffer size in multiples of
sample-data-blocks. A sample-data-block occupies approximately 4 KB. The
specified minimum value is compared with the initial buffer size that is
calculated by the sampling facility. The greater value is then used as the initial
size when the sampling facility is started.

3. Optional: Specify the `chcpumf` command with the `-x` parameter to set the
maximum buffer size.

Example:

```
# chcpumf -x 1000
```
The value that you specify with -x is the maximum buffer size in multiples of sample-data-blocks. A sample-data-block occupies approximately 4 KB. The specified maximum is the upper limit to which the sampling facility can adjust the buffer.

**Example**

**Tips:**
- You can specify both, the minimum and the maximum buffer size with a single command.
- Use the -V parameter to display the minimum and maximum buffer settings that apply as a result of the command.

**Example:** To change the minimum buffer size to 500 times the size of a sample-data-block and the maximum buffer size to 1000 times the size of a sample-data-block, issue:

```bash
# chcpumf -V -m 500 -x 1000
Sampling buffer sizes:
  Minimum: 500 sample-data-blocks
  Maximum: 1000 sample-data-blocks
```

**Obtaining debug information**

You can obtain version information for the CPU-measurement counter facility and check which counter sets are authorized on your LPAR.

**Before you begin**

If you call magic syscall functions with a method other than through the procfs, you might need to activate them first. For more information about the magic syscall functions, see "Using the magic syscall feature" on page 49.

**Procedure**

Perform these steps to obtain debug information:

1. Use the magic syscall function with character p to trigger a kernel message with information about the CPU-measurement counter facility.
   For example, trigger the message from procfs:
   ```bash
   # echo p > /proc/sysrq-trigger
   ```

2. Find the message by issuing the `dmesg` command and looking for output lines that include CPUM_CF.

   **Example:**
   ```bash
   CPU[0] CPUM_CF: ver=1.2 A=0000c E=0008 C=0000
   ```

   **Note:** The message is specific to the particular processor that processed the magic syscall. However, the scope of the version (ver=) and authorization (A=) information is the LPAR and can be read from the message for any processor in the LPAR. The values for E= (enabled) and C= (activated) can differ among processors.

3. Obtain the version of the CPU-measurement counter facility by reading the value of the ver= parameter in the message.
4. Check whether counter sets are authorized for the LPAR by interpreting the value of the $A=$ parameter in the message.

The value is a 4-digit hexadecimal number that represents the sums of these values for the individual counter sets:

- 0001: Extended counter set.
- 0002: Basic counter set.
- 0004: Problem-state counter set.
- 0008: Crypto-activity counter set.

**Examples:**

- $A=0000$ means that none of the counter set are authorized.
- $A=000c$ means that the Problem-state counter set and the Crypto-activity counter set are authorized.
- $A=000f$ means that all four counter sets are authorized.

**More information:** For more details, see z/Architecture The Load-Program-Parameter and the CPU-Measurement Facilities, SA23-2260.

**Example**

This example shows how to trigger the message from procfs and how to find the message:

```
# echo p > /proc/sysrq-trigger
# dmesg | grep CPUM_CF
CPU[0] CPUM_CF: ver=1.2 A=000c E=0008 C=0000
```

In the message, ver=1.2 means version 1.2 of the IBM Z CPU-measurement counter facility.

Because $0x000c = 0x0004 + 0x0008$, the $A=000c$ of the example means that the Problem-state counter set and the Crypto-activity counter set are authorized for the LPAR.

**cpu0 only:** $E=0008$ means that only the Crypto-activity counter set is enabled, and the $C=0000$ means that neither of the counter sets are activated.
Part 9. Diagnostics and troubleshooting

Chapter 48. Logging I/O subchannel status information ................................................. 473
Chapter 49. Control program identification ................. 475
   Specifying a system name ................. 475
   Specifying a sysplex name ................. 475
   Specifying a system type .................. 476
   Specifying the system level ............... 476
   Sending system data to the SE. ........... 477

Chapter 48. Logging I/O subchannel status information.

Chapter 49. Control program identification ................. 475
   Specifying a system name ................. 475
   Specifying a sysplex name ................. 475
   Specifying a system type .................. 476
   Specifying the system level ............... 476
   Sending system data to the SE. ........... 477

Chapter 50. Activating automatic problem reporting ................. 479
   Setting up the Call Home support .......... 479
   Activating the Call Home support .......... 479

Chapter 51. Check whether the Linux instance can be a hypervisor ................. 481

Chapter 52. Avoiding common pitfalls ......................... 483
   Ensuring correct channel path status ...... 483
   Determining channel path usage .......... 484
   Configuring LPAR I/O devices .............. 484
   Using cio_ignore ......................... 484
   Excessive guest swapping ................ 484
   Including service levels of the hardware and the hypervisor ......... 485
   Booting stops with disabled wait state ... 485
   Preparing for dump-on-panic .............. 485
   Multipath failover causes kernel panic ... 486

These resources are useful when diagnosing and solving problems for Red Hat Enterprise Linux 7.5.

Newest version

You can find the newest version of this book at
www.ibm.com/support/knowledgcenter/linuxonibm/liaaf/lnz_r_distlibs.html

Restrictions

For prerequisites and restrictions see the IBM Z architecture specific information in the Red Hat Enterprise Linux 7.5 release notes at
https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux

When reporting a problem to IBM support, you might be asked to supply a kernel dump. See Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711 for information about how to create dumps.
Chapter 48. Logging I/O subchannel status information

When investigating I/O subchannels, support specialists might request operation status information for the subchannel.

About this task

The channel subsystem offers a logging facility that creates a set of log entries with such information. From Linux, you can trigger this logging facility through sysfs.

The log entries are available through the SE Console Actions Work Area with the View Console Logs function. The entries differ dependent on the device and model that is connected to the subchannel. On the SE, the entries are listed with a prefix that identifies the model. The content of the entries is intended for support specialists.

Procedure

To create a log entry, issue a command of this form:

```bash
# echo 1 > /sys/devices/css0/<subchannel-bus-id>/logging
```

where `<subchannel-bus-id>` is the bus ID of the I/O subchannel that corresponds to the I/O device for which you want to create a log entry.

To find out how your I/O devices map to subchannels you can use, for example, the `lscss` command.

Example

In this example, first the subchannel for an I/O device with bus ID `0.0.3d07` is identified, then logging is initiated.

```
# lscss -d 0.0.3d07
Device  Subchan  DevType CU Type Use  PIM  PAM  POM  CHPIDs
-----------------------------------------------
0.0.3d07 0.0.000c 1732/01 1731/01  80  80  ff  05000000 00000000
# echo 1 > /sys/devices/css0/0.0.000c/logging
```
Chapter 49. Control program identification

For Linux in LPAR mode, you can provide data about the Linux instance to the control program identification (CPI) feature.

The data is used, for example, to represent the Linux instance on the HMC or SE.

You provide data to the CPI feature in two steps:
1. Write values for one or more of the following items to specific sysfs attributes in /sys/firmware/cpi:
   - The name of the Linux instance
   - The sysplex name (if applicable)
   - The operating system type
   - The operating system level
2. Transfer the data to the SE, see “Sending system data to the SE” on page 477.

Red Hat Enterprise Linux 7.5 provides an init script, /etc/init.d/cpi, to set the system and sysplex name automatically during system boot.

Specifying a system name

Use the system_name attribute in the /sys/firmware/cpi directory in sysfs to specify a system name for your Linux instance.

About this task

The system name is a string that consists of up to eight characters of the following set: A-Z, 0-9, $, @, #, and blank.

Example

```
# echo LPAR12 > /sys/firmware/cpi/system_name
```

What to do next

To make the setting take effect, transfer the data to the SE (see “Sending system data to the SE” on page 477).

Specifying a sysplex name

Use the sysplex_name attribute in the /sys/firmware/cpi directory in sysfs to specify a sysplex name.

About this task

The sysplex name is a string that consists of up to eight characters of the following set: A-Z, 0-9, $, @, #, and blank.
Example

```
# echo SYSPLEX1 > /sys/firmware/cpi/sysplex_name
```

What to do next

To make the setting take effect, transfer the data to the SE (see “Sending system data to the SE” on page 477).

Specifying a system type

Linux uses the `/sys/firmware/cpi/system_type` sysfs attribute to identify itself as a Linux instance.

About this task

Unless your distribution sets this value for you, write `LINUX` to the attribute.

Example

```
# cat /sys/firmware/cpi/system_type
```
```
**
```
# echo LINUX > /sys/firmware/cpi/system_type
```

What to do next

To make the setting take effect, transfer the data to the SE (see “Sending system data to the SE” on page 477).

Specifying the system level

Linux uses the `/sys/firmware/cpi/system_level` sysfs attribute for the kernel version.

About this task

The value has this format:

```
0x0000000000031200
```

where:

- `<aa>` are two digits for the major version of the kernel.
- `<bb>` are two digits for the minor version of the kernel.
- `<cc>` are two digits for the stable version of the kernel.

Example

Linux kernel 3.12 displays as

```
# cat /sys/firmware/cpi/system_level
0x0000000000031200
```
What to do next

To make the setting take effect, transfer the data to the SE (see "Sending system data to the SE").

---

**Sending system data to the SE**

Use the set attribute in the `/sys/firmware/cpi` directory in sysfs to send data to the service element.

**About this task**

To send the data in attributes `sysplex_name`, `system_level`, `system_name`, and `system_type` to the SE, write an arbitrary string to the set attribute.

**Example**

```bash
# echo 1 > /sys/firmware/cpi/set
```
Chapter 50. Activating automatic problem reporting

You can activate automatic problem reporting for situations where Linux experiences a kernel panic.

Before you begin

- The Linux instance must run in an LPAR.
- You need a hardware support agreement with IBM to report problems to RETAIN.

About this task

Linux uses the Call Home function to send automatically collected problem data to the IBM service organization through the Service Element. Hence a system crash automatically leads to a new Problem Management Record (PMR) which can be processed by IBM service.

Setting up the Call Home support

To set up the Call Home support, load the sclp_async module with the `modprobe` command.

About this task

There are no module parameters for sclp_async.

Procedure

Load the sclp_async module with the `modprobe` command to ensure that any other required modules are loaded in the correct order:

```
# modprobe sclp_async
```

Activating the Call Home support

When the sclp_async module is loaded, you can control it through the sysctl interface or procfs.

Procedure

To activate the support, set the `callhome` attribute to 1. To deactivate the support, set the `callhome` attribute to 0. Issue a command of this form:

```
# echo <flag> > /proc/sys/kernel/callhome
```

This is equivalent to:

```
# sysctl -w kernel.callhome=<flag>
```

To persistently enable the Call Home support across reboots, add "kernel.callhome=1" to the `/etc/sysctl.conf` file. The kernel module must be
loaded before the sysctl.conf file is processed. Linux cannot check whether Call Home is supported by the hardware.

**Examples**

- To activate the Call Home support, issue:

  ```
  # echo 1 > /proc/sys/kernel/callhome
  ```

- To deactivate the Call Home support, issue:

  ```
  # echo 0 > /proc/sys/kernel/callhome
  ```
Chapter 51. Check whether the Linux instance can be a hypervisor

An instance of Linux on Z must have the SIE (Start Interpretive Execution) capability to be able to act as a hypervisor, such as a KVM host.

Procedure

1. Issue the following command to find out whether you can operate your Linux instance as a hypervisor.

```
# cat /proc/cpuinfo
vendor_id : IBM/S390
# processors : 1
bogomips per cpu: 14367.00
features : esan3 zarch stfle msa ldisp eimm dfp edat etf3eh
highgprs sie
cache0 : level=1 type=Data scope=Private size=128K
...
```

2. Examine the features line in the command output. If the list of features includes s1e, the Linux instance can be a hypervisor. The Linux instance of the example can be a hypervisor.
Chapter 52. Avoiding common pitfalls

Common problems and how to avoid them.

Ensuring correct channel path status

Ensure that you have varied the path offline before you perform a planned task on it.

Tasks that require the channel path to be offline include:
- Pulling out or plugging in a cable on a path.
- Configuring a path off or on at the SE.

To vary the path offline, issue a command of the form:

```
# chchp -v 0 <chpid>
```

where `<chpid>` is the channel path ID.

After the operation completed and the path is available again, vary the path online by using a command of the form:

```
# chchp -v 1 <chpid>
```

Alternatively, you can write `on` or `off` to the channel path status attribute in sysfs to vary the path online or offline.

```
# echo on|off > /sys/devices/css0/chp0.<chpid>/status
```

An unplanned change in path availability can occur due to, for example, unplanned cable pulls or a temporary path malfunction. Then, the PIM/PAM/POM values (as obtained through `lscss`) might not be as expected. To update the PIM/PAM/POM values, vary one of the paths that lead to the affected devices.

Example:

```
# chchp -v 0 0.12
# chchp -v 1 0.12
```

Rationale: Linux does not always receive a notification (machine check) when the status of a path changes (especially for a path that comes online again). To make sure Linux has up-to-date information about the usable paths, path verification is triggered through the Linux vary operation.
Determining channel path usage

To determine the usage of a specific channel path on LPAR, for example, to check whether traffic is distributed evenly over all channel paths, use the channel path measurement facility.

See “Channel path measurement” on page 14 for details.

Configuring LPAR I/O devices

A Linux LPAR should contain only those I/O devices that it uses.

Limit the I/O devices by:
- Adding only the needed devices to the IOCDS
- Using the cio_ignore kernel parameter to ignore all devices that are not currently in use by this LPAR.

If more devices are needed later, they can be dynamically removed from the list of devices to be ignored. Use the cioignore kernel parameter or the /proc/cio_ignore dynamic control to remove devices, see “cio_ignore - List devices to be ignored” on page 652 and “Changing the exclusion list” on page 653.

Rationale: Numerous unused devices can cause:
- Unnecessary high memory usage due to allocation of device structures.
- Unnecessary high load on status changes because hot-plug handling must be done for every device found.

Using cio_ignore

With cioignore, essential devices might be hidden.

For example, if Linux does not boot under z/VM and does not show any message except:
HCPGIR450W CP entered; disabled wait PSW 00020001 80000000 00000000 00144D7A

Check if cioignore is used and verify that the console device, which is typically device number 0.0.0009, is not ignored.

Excessive guest swapping

Avoid excessive guest swapping by using the timed page pool size and the static page pool size attributes.

An instance of Linux on z/VM might be swapping and stalling. Setting the timed page pool size and the static page pool size to zero might solve the problem:

```
# echo 0 > /proc/sys/vm/cmm_timed_pages
# echo 0 > /proc/sys/vm/cmm_pages
```

If you see a temporary relief, the guest does not have enough memory. Try increasing the guest memory.

If the problem persists, z/VM might be out of memory.
If you are using cooperative memory management (CMM), unload the cooperative memory management module:

```
# modprobe -r cmm
```

See Chapter 38, “Cooperative memory management,” on page 421 for more details about CMM.

---

### Including service levels of the hardware and the hypervisor

The service levels of the different hardware cards, the LPAR level, and the z/VM service level are valuable information for problem analysis.

If possible, include this information with any problem you report to IBM service.

A `/proc` interface that provides a list of service levels is available. To see the service levels issue:

```
# cat /proc/service_levels
```

Example for a z/VM system with a QETH adapter:

```
# cat /proc/service_levels
VM: z/VM Version 5 Release 2.0, service level 0801 (64-bit)
qeth: 0.0.5f0 firmware level 087d
```

---

### Booting stops with disabled wait state

An automatic processor type check might stop the boot process with a disabled wait PSW.

On Red Hat Enterprise Linux 7.5, a processor type check is automatically run at every kernel startup. If the check determines that Red Hat Enterprise Linux 7.5 is not compatible with the hardware, it stops the boot process with a disabled wait PSW with an address of zero.

If this problem occurs, ensure that you are running Red Hat Enterprise Linux 7.5 on supported hardware. See the Red Hat Enterprise Linux 7.5 release notes at [https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux](https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux)

---

### Preparing for dump-on-panic

You might want to consider setting up your system to automatically create a memory dump after a kernel panic.

Configuring and using dump-on-panic has the following advantages:

- You have a memory dump disk that is prepared ahead of time.
- You do not have to reproduce the problem since a memory dump will be triggered automatically immediately after the failure.

Multipath failover causes kernel panic

In a multipath setup where SCSI disks are attached over multiple paths, failover might trigger a kernel panic.

To remedy this, try increasing the value for `fast_fail` to, for example, 5 seconds and `dev_loss_tmo` to, for example, 120 seconds.
## Part 10. Reference

### Chapter 53. Commands for Linux on Z

<table>
<thead>
<tr>
<th>Command</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic command options</td>
<td>489</td>
</tr>
<tr>
<td>chccwdev - Set CCW device attributes</td>
<td>490</td>
</tr>
<tr>
<td>chcp - Change channel path status</td>
<td>492</td>
</tr>
<tr>
<td>chmem - Set memory online or offline</td>
<td>494</td>
</tr>
<tr>
<td>chcpumf - Set limits for the CPU measurement sampling facility buffer</td>
<td>496</td>
</tr>
<tr>
<td>chreipl - Modify the re-IPL configuration</td>
<td>497</td>
</tr>
<tr>
<td>chshut - Control the system shutdown actions</td>
<td>501</td>
</tr>
<tr>
<td>chzcrypt - Modify the cryptographic configuration</td>
<td>503</td>
</tr>
<tr>
<td>cmsfs-fuse - Mount a z/VM CMS file system</td>
<td>505</td>
</tr>
<tr>
<td>cpuplugd - Control CPUs and memory</td>
<td>510</td>
</tr>
<tr>
<td>dasdfmt - Format a DASD</td>
<td>520</td>
</tr>
<tr>
<td>dasdstat - Display DASD performance statistics</td>
<td>525</td>
</tr>
<tr>
<td>dasdview - Display DASD structure</td>
<td>528</td>
</tr>
<tr>
<td>fdasd - Partition a DASD</td>
<td>539</td>
</tr>
<tr>
<td>hmcdrvfs - Mount a FUSE file system for remote access to media in the HMC media drive</td>
<td>547</td>
</tr>
<tr>
<td>hytop - Display hypervisor performance data</td>
<td>551</td>
</tr>
<tr>
<td>lschp - List channel paths</td>
<td>562</td>
</tr>
<tr>
<td>lscpumf - Display information about the CPU-measurement facilities</td>
<td>564</td>
</tr>
<tr>
<td>lsccs - List subchannels</td>
<td>567</td>
</tr>
<tr>
<td>lsdasd - List DASD devices</td>
<td>571</td>
</tr>
<tr>
<td>lshmc - List media contents in the HMC media drive</td>
<td>575</td>
</tr>
<tr>
<td>lslns - Discover LUNs, or show encryption state of attached LUNs</td>
<td>576</td>
</tr>
<tr>
<td>lsmem - Show online status information about memory blocks</td>
<td>579</td>
</tr>
<tr>
<td>lsqeth - List qeth-based network devices</td>
<td>581</td>
</tr>
<tr>
<td>lsreipl - List IPL and re-IPL settings</td>
<td>583</td>
</tr>
<tr>
<td>lsscm - List storage-class memory increments</td>
<td>584</td>
</tr>
<tr>
<td>lsshut - List the current system shutdown actions</td>
<td>586</td>
</tr>
<tr>
<td>lstape - List tape devices</td>
<td>587</td>
</tr>
<tr>
<td>lszcrypt - Display cryptographic devices</td>
<td>591</td>
</tr>
<tr>
<td>lszfp - List zfp devices</td>
<td>595</td>
</tr>
<tr>
<td>mon_fssstatd - Monitor z/VM guest file system size</td>
<td>597</td>
</tr>
<tr>
<td>mon_procmd - Monitor Linux on z/VM</td>
<td>601</td>
</tr>
<tr>
<td>qetharp - Query and purge OSA and HiperSockets ARP data</td>
<td>608</td>
</tr>
<tr>
<td>qethconf - Configure qeth devices</td>
<td>610</td>
</tr>
<tr>
<td>qethqoat - Query OSA address table</td>
<td>613</td>
</tr>
<tr>
<td>scsi_logging_level - Set and get the SCSI logging level</td>
<td>616</td>
</tr>
<tr>
<td>tape390_crypt - Manage tape encryption</td>
<td>619</td>
</tr>
<tr>
<td>tape390_display - Display messages on tape devices and load tapes</td>
<td>623</td>
</tr>
<tr>
<td>tunedasd - Adjust low-level DASD settings</td>
<td>625</td>
</tr>
<tr>
<td>vmcp - Send CP commands to the z/VM hypervisor</td>
<td>629</td>
</tr>
<tr>
<td>vmsr - Work with z/VM spool file queues</td>
<td>631</td>
</tr>
<tr>
<td>zdfs - Mount a z/OS DASD</td>
<td>640</td>
</tr>
<tr>
<td>zkey - Generate, validate, and re-encrypt secure AES keys</td>
<td>645</td>
</tr>
<tr>
<td>znetconf - List and configure network devices</td>
<td>648</td>
</tr>
</tbody>
</table>

### Chapter 54. Selected kernel parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>cio_ignore - List devices to be ignored</td>
<td>651</td>
</tr>
<tr>
<td>cmma - Reduce hypervisor paging I/O overhead</td>
<td>652</td>
</tr>
<tr>
<td>fps - Run Linux in FIPS mode</td>
<td>657</td>
</tr>
<tr>
<td>maxcpus - Limit the number of CPUs Linux can use at IPL</td>
<td>658</td>
</tr>
<tr>
<td>nosmt - Disable simultaneous multithreading</td>
<td>659</td>
</tr>
<tr>
<td>possible_cpus - Limit the number of CPUs Linux can use</td>
<td>660</td>
</tr>
<tr>
<td>ramdisk_size - Specify the ramdisk size</td>
<td>661</td>
</tr>
<tr>
<td>root - Mount the root file system read-only</td>
<td>662</td>
</tr>
<tr>
<td>smt - Reduce the number of threads per core</td>
<td>663</td>
</tr>
<tr>
<td>vdsio - Optimize system call performance</td>
<td>664</td>
</tr>
<tr>
<td>vmhalt - Specify CP command to run after a system halt</td>
<td>665</td>
</tr>
<tr>
<td>vmppanic - Specify CP command to run after a kernel panic</td>
<td>666</td>
</tr>
<tr>
<td>vmpoff - Specify CP command to run after a power off</td>
<td>667</td>
</tr>
<tr>
<td>vmreboot - Specify CP command to run on reboot</td>
<td>668</td>
</tr>
</tbody>
</table>

### Chapter 55. Linux diagnose code use

Use these commands, kernel parameters, kernel options to configure Linux on Z. Be aware of the z/VM DIAG calls required by Red Hat Enterprise Linux 7.5.

## Newest version

You can find the newest version of this book at


## Restrictions

For prerequisites and restrictions see the IBM Z architecture specific information in the Red Hat Enterprise Linux 7.5 release notes at

Chapter 53. Commands for Linux on Z

You can use z/Architecture specific commands to configure and work with the Red Hat Enterprise Linux 7.5 for z Systems device drivers and features.

Some commands come with an init script or a configuration file or both. Init scripts are installed in /etc/init.d/. You can extract any missing files from the etc subdirectory in the s390utils RPM.

Commands described elsewhere

- For the zipl command, see Chapter 5, “Initial program loader for IBM Z - zipl,” on page 55.
- For commands and tools that are related to creating and analyzing system dumps, see Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711.
- For commands related to terminal access over IUCV connections, see How to Set up a Terminal Server Environment on z/VM, SC34-2596.
- The icainfo and icastats commands are provided with the libica package and described in libica Programmer’s Reference, SC34-2602.

Generic command options

For simplicity, common command options are omitted from some of the syntax diagrams.

- **-h** or **--help**
  to display help information for the command.

- **--version**
  to display version information for the command.

The syntax for these options is:

```
Common command options

<command> Other command options

-h
--help
--version

```

where **command** can be any of the Linux on Z commands.

chccwdev - Set CCW device attributes

Use the `chccwdev` command to set attributes for CCW devices and to set CCW devices online or offline.

Use "znetconf - List and configure network devices" on page 648 to work with CCW_GROUP devices. For more information about CCW devices and CCW group devices, see "Device categories" on page 7.

The `chccwdev` command uses cio_settle before it changes anything, to ensure that sysfs reflects the latest device status information, and includes newly available devices.

chccwdev syntax

```
chccwdev -e -d -s -f
-a <name>=<value>
-device_bus_id=<device_bus_id>
-from_device_bus_id=-<to_device_bus_id>
```

Where:

- **-e or --online**
  sets the device online.

- **-d or --offline**
  sets the device offline.

- **-s or --safeoffline**
  waits until all outstanding I/O requests complete, and then tries to set the device offline. Valid for DASDs only.

- **-f or --forceonline**
  forces a boxed device online, if this action is supported by the device driver.

- **-a <name>=<value> or --attribute <name>=<value>**
  sets the <name> attribute to <value>.

  The available attributes depend on the device type. See the chapter for your device for details about the applicable attributes and values.

  Setting the online attribute has the same effect as using the -e or -d options.

- **<device_bus_id>**
  identifies a device. Device bus-IDs are of the form 0.<n>.<devno>, where <n> is a subchannel set ID and <devno> is a device number. Input is converted to lowercase.
<from_device_bus_id>-<to_device_bus_id>
identifies a range of devices. If not all devices in the range exist, the command
is limited to the existing ones. If you specify a range with no existing devices,
you get an error message.

-h or --help
displays help information for the command. To view the man page, enter man
chccwdev.

-v or --version
displays version information for the command.

Examples

- To set a CCW device 0.0.b100 online issue:

```bash
# chccwdev -e 0.0.b100
```

- Alternatively, use -a to set a CCW device 0.0.b100 online. Issue:

```bash
# chccwdev -a online=1 0.0.b100
```

- To set all CCW devices in the range 0.0.b200 through 0.0.b2ff online, issue:

```bash
# chccwdev -e 0.0.b200-0.0.b2ff
```

- To set a CCW device 0.0.b100 and all CCW devices in the range 0.0.b200 through
  0.0.b2ff offline, issue:

```bash
# chccwdev -d 0.0.b100,0.0.b200-0.0.b2ff
```

- To set several CCW devices in different ranges and different subchannel sets
  offline, issue:

```bash
# chccwdev -a online=0 0.0.1000-0.0.1100,0.1.7000-0.1.7010,0.0.1234,0.1.4321
```

- To set devices with bus ID 0.0.0192, and 0.0.0195 through 0.0.0198 offline after
  completing all outstanding I/O requests:

```bash
# chccwdev -s 0.0.0192,0.0.0195-0.0.0198
```

If an outstanding I/O request is blocked, the command might wait forever.
Reasons for blocked I/O requests include reserved devices that can be released
or disconnected devices that can be reconnected.

1. Try to resolve the problem that blocks the I/O request and wait for the
   command to complete.
2. If you cannot resolve the problem, issue chccwdev -d to cancel the
   outstanding I/O requests. The data is lost.

- To set an ECKD DASD 0.0.b100 online and to enable extended error reporting
  and logging issue:

```bash
# chccwdev -e -a eer_enabled=1 -a erplog=1 0.0.b100
```
chchp - Change channel path status

Use the `chchp` command to set channel paths online or offline.

The actions are equivalent to performing a Configure Channel Path Off or Configure Channel Path On operation on the Hardware Management Console.

The channel path status that results from a configure operation is persistent across IPLs.

**Note:** Changing the configuration state of an I/O channel path might affect the availability of I/O devices. It can also trigger associated functions (such as channel-path verification or device scanning) which in turn can result in a temporary increase in processor, memory, and I/O load.

**chchp syntax**

```
chchp [-c [<value> | --configure <value>]] [-v [<value> | --vary <value>]] [-a <key>=<value>]
```

Where:

- `-c <value>` or `--configure <value>`
  
  sets the device to configured (1) or standby (0).

  **Note:** Setting the configured state to standby can stop running I/O operations.

- `-v <value>` or `--vary <value>`

  changes the logical channel-path state to online (1) or offline (0).

  **Note:** Setting the logical state to offline can stop running I/O operations.

- `-a <key>=<value>` or `--attribute <key>=<value>`

  changes the channel-path sysfs attribute `<key>` to `<value>`. The `<key>` can be the name of any available channel-path sysfs attribute (that is, configure or status). `<value>` can take any valid value that can be written to the attribute (for example, 0 or offline). Using `-a` is a generic way of writing to the corresponding sysfs attribute. It is intended for cases where sysfs attributes or attribute values are available in the kernel but not in chchp.

- `0.<id>` and `0.<id> - 0.<id>`

  where `<id>` is a hexadecimal, two-digit, lowercase identifier for the channel path. An operation can be performed on more than one channel path by specifying multiple identifiers as a comma-separated list, or a range, or a combination of both.

- `--version`

  displays the version number of chchp and exits.
-h or --help

displays a short help text, then exits.

Examples

- To set channel path 0.19 into standby state issue:

```bash
# chchp -a configure=0 0.19
```

- To set the channel path with the channel path ID 0.40 to the standby state, write 0 to the configure file with the chchp command:

```bash
# chchp --configure 0 0.40
Configure standby 0.40... done.
```

- To set a channel-path to the configured state, write 1 to the configure file with the chchp command:

```bash
# chchp --configure 1 0.40
Configure online 0.40... done.
```

- To set channel-paths 0.65 to 0.6f to the configured state issue:

```bash
# chchp -c 1 0.65-0.6f
```

- To set channel-paths 0.12, 0.7f and 0.17 to 0.20 to the logical offline state issue:

```bash
# chchp -v 0 0.12,0.7f,0.17-0.20
```
chmem

chmem - Set memory online or offline

Use the chmem command to set a particular size or range of memory online or offline.

Setting memory online can fail if the hypervisor does not have enough memory left, for example because memory was overcommitted. Setting memory offline can fail if Linux cannot free the memory. If only part of the requested memory can be set online or offline, a message informs you how much memory was set online or offline instead of the requested amount.

chmem syntax

```
chmem [-e | --enable <size>] [-d | --disable <start>-<end>]
```

Where:

- **-e or --enable**
  sets the specified memory online.

- **-d or --disable**
  sets the specified memory offline.

- **<size>**
  specifies an amount of memory to be set online or offline. A numeric value without a unit or a numeric value immediately followed by m or M is interpreted as MB (1024 x 1024 bytes). A numeric value immediately followed by g or G is interpreted as GB (1024 x 1024 x 1024 bytes).

  The size must be aligned to the memory block size, as shown in the output of the lsmem command.

- **<start>-<end>**
  specifies a memory range to be set online or offline. `<start>` is the hexadecimal address of the first byte and `<end>` is the hexadecimal address of the last byte in the memory range.

  The range must be aligned to the memory block size, as shown in the output of the lsmem command.

- **-v or --version**
  displays the version number of chmem, then exits.

- **-h or --help**
  displays a short help text, then exits. To view the man page, enter `man chmem`.

Examples

- This command requests 1024 MB of memory to be set online.
  
  ```
  # chmem --enable 1024
  ```

- This command requests 2 GB of memory to be set online.
  
  ```
  # chmem --enable 2g
  ```
• This command requests the memory range that starts with 0x00000000e4000000 and ends with 0x00000000f3ffffff to be set offline.

```
# chmem --disable 0x00000000e4000000-0x00000000f3ffffff
```
chcpumf

chcpumf - Set limits for the CPU measurement sampling facility buffer

Use the chcpumf command to set limits for the CPU measurement sampling facility buffer.

The sampling facility is designed for autonomous buffer management, and you do not usually need to intervene. However, you might want to change the minimum or maximum size, for example, for one of the following reasons:

- There are considerable resource constraints on your system, and the sampling facility stops because it tries to allocate more buffer space than is available.
- As an expert user of perf and the sampling facility, you want to explore results with particular buffer settings.

chcpumf syntax

```
chcpumf [-m <min_sdb>] [-V] [-x <max_sdb>]
```

where:

- `-m <min_sdb>` or `--min <min_sdb>`
  specifies the minimum sampling facility buffer size in sample-data-blocks. A sample-data-block occupies approximately 4 KB. The sampling facility starts with this buffer size if it exceeds the initial buffer size that is calculated by the sampling facility.

- `-x <max_sdb>` or `--max <max_sdb>`
  specifies the maximum sampling facility buffer size in sample-data-blocks. A sample-data-block occupies approximately 4 KB. While it is running, the sampling facility dynamically adjusts the buffer size to a suitable value, but cannot exceed this limit.

- `-V` or `--verbose`
  displays the buffer size settings after the changes.

- `-v` or `--version`
  displays the version number of chcpumf and exits.

- `-h` or `--help`
  displays out a short help text, then exits. To view the man page, enter `man chcpumf`.

Example

To change the minimum buffer size to 500 times the size of a sample-data-block and the maximum buffer size to 1000 times the size of a sample-data-block, issue:

```
# chcpumf -V -m 500 -x 1000
Sampling buffer sizes:
Minimum: 500 sample-data-blocks
Maximum: 1000 sample-data-blocks
```
chreipl - Modify the re-IPL configuration

Use the chreipl tool to modify the re-IPL configuration for Linux on Z.

You can configure a particular device as the re-IPL device. For zipl boot menu configurations, you can set the boot menu entry to be used for the next re-IPL. You can also specify additional kernel parameters for the next re-IPL.

chreipl syntax

```
chreipl [ccw|fcp] -L <parm> [1]
  <device_bus_id> [fcp] <device_bus_id> <wwpn> <lun>-b <n> [2]
  nss <name> [nss] -p < parms >
```

Notes:
1  You can specify the <device_bus_id>, <wwpn>, and <lun> in any order if you use the corresponding command options.
2  -L can be used if the device node or directory maps to a DASD; -b can be used if the device node or directory maps to a SCSI disk.

Where:
- `<device_bus_id>` or `-d <device_bus_id>` or `--device <device_bus_id>`
  specifies the device bus-ID of a CCW re-IPL device or of the FCP device through with a SCSI re-IPL device is attached.
- `<wwpn>` or `-w <wwpn>` or `--wwpn <wwpn>`
  specifies the worldwide port name (WWPN) of a SCSI re-IPL device.
- `<lun>` or `-l <lun>` or `--lun <lun>`
  specifies the logical unit number (LUN) of a SCSI re-IPL device.
- `<node>`
  specifies a device node of a DASD, SCSI, or logical device mapper re-IPL device. For more information about logical boot devices, see “Preparing a logical device as a boot device” on page 62.
- `<dir>`
  specifies a directory in the Linux file system on the re-IPL device.
chreipl

**nss**
declares that the following parameters refer to a z/VM named saved system (NSS).

<name> or -n <name> or --name <name>
specifies the name of an NSS as defined on the z/VM system.

-L <parameter> or --loadparm <parameter>
specifies the entry in the boot menu to be used for the next reboot. This parameter applies only if the re-IPL device is a DASD with a zipl boot menu configuration.

Omitting this parameter eliminates an existing selection in the boot configuration. Depending on your boot menu configuration, a zipl interactive boot menu might be displayed during the re-IPL process or the default configuration is used. See [“Example for a DASD menu configuration on z/VM” on page 86](#), [“Example for a DASD menu configuration (LPAR)” on page 91](#), and [“Menu configurations” on page 77](#) for details.

-b <n> or --bootprog <n>
specifies the entry in the boot menu to be used for the next reboot. This parameter applies only if the re-IPL device is a SCSI disk with a zipl boot menu configuration.

Omitting this parameter eliminates an existing selection in the boot configuration and the default boot configuration is used.

-p or --bootparms
specifies boot parameters for the next reboot. The boot parameters, which typically are kernel parameters, are appended to the kernel parameter line in the boot configuration. The number of characters you can specify depends on your environment and re-IPL device as shown in Table 59.

Table 59. Maximum characters for additional kernel parameters

<table>
<thead>
<tr>
<th>Virtual hardware where Linux runs</th>
<th>DASD re-IPL device</th>
<th>SCSI re-IPL device</th>
<th>NSS re-IPL device</th>
</tr>
</thead>
<tbody>
<tr>
<td>z/VM guest virtual machine</td>
<td>64</td>
<td>3452</td>
<td>56</td>
</tr>
<tr>
<td>LPAR</td>
<td>none</td>
<td>3452</td>
<td>n/a</td>
</tr>
</tbody>
</table>

If you omit this parameter, the existing boot parameters in the next boot configuration are used without any changes.

-h or --help

displays help information for the command. To view the man page, enter `man chreipl`.

-v or --version

displays version information.

For disk-type re-IPL devices, the command accepts but does not require an initial statement:

ccw

declares that the following parameters refer to a DASD re-IPL device.

fcp

declares that the following parameters refer to a SCSI re-IPL device.

node

declares that the following parameters refer to a disk re-IPL device that is
identified by a device node or by a directory in the Linux file system on that device. The disk device can be a DASD or a SCSI disk.

**Examples**

These examples illustrate common uses for `chreipl`.

- The following commands all configure the same DASD as the re-IPL device, assuming that the device bus-ID of the DASD is `0.0.7e78`, that the standard device node is `/dev/dasdc`, that udev creates an alternative device node `/dev/disk/by-path/ccw-0.0.7e78`, that `/mnt/boot` is located on the Linux file system in a partition of the DASD.
  - Using the bus ID:
    ```
    # chreipl 0.0.7e78
    ```
  - Using the bus ID and the optional `ccw` statement:
    ```
    # chreipl ccw 0.0.7e78
    ```
  - Using the bus ID, the optional statement and the optional `--device` keyword:
    ```
    # chreipl ccw --device 0.0.7e78
    ```
  - Using the standard device node:
    ```
    # chreipl /dev/dasdc
    ```
  - Using the udev-created device node:
    ```
    # chreipl /dev/disk/by-path/ccw-0.0.7e78
    ```
  - Using a directory within the file system on the DASD:
    ```
    # chreipl /mnt/boot
    ```

- The following commands all configure the same SCSI disk as the re-IPL device, assuming that the device bus-ID of the FCP device through which the device is attached is `0.0.1700`, the WWPN of the storage server is `0x500507630300c562`, and the LUN is `0x401040b300000000`. Further it is assumed that the standard device node is `/dev/sdb`, that udev creates an alternative device node `/dev/disk/by-id/scsi-36005076303fffc5620000000000010b4`, and that `/mnt/fcpboot` is located on the Linux file system in a partition of the SCSI disk.
  - Using bus ID, WWPN, and LUN:
    ```
    # chreipl 0.0.1700 0x500507630300c562 0x401040b300000000
    ```
  - Using bus ID, WWPN, and LUN with the optional `fcp` statement:
    ```
    # chreipl fcp 0.0.1700 0x500507630300c562 0x401040b300000000
    ```
  - Using bus ID, WWPN, LUN, the optional statement, and keywords for the parameters. When you use the keywords, the parameters can be specified in any order:
    ```
    # chreipl fcp --wwpn 0x500507630300c562 -d 0.0.1700 --lun 0x401040b300000000
    ```
  - Using the standard device node:
- Using the udev-created device node:

```
# chreipl /dev/disk/by-id/scsi-36005076303ffcc5620000000000010b4
```

- Using a directory within the file system on the SCSI disk:

```
# chreipl /mnt/fcpboot
```

- To configure a DASD with bus ID 0.0.7e78 as the re-IPL device, using the first entry of the zipl boot menu:

```
# chreipl 0.0.7e78 -L 1
Re-IPL type: ccw
Device: 0.0.7e78
Loadparm: "1"
Bootparms: ""
```

- The following examples configures a DASD with bus ID 0.0.7e78 as the re-IPL device. To also display all kernel messages on the console, add `ignore_loglevel` to the existing kernel parameters in the boot configuration:

```
# chreipl 0.0.7e78 -p "ignore_loglevel"
Re-IPL type: ccw
Device: 0.0.7e78
Loadparm: ""
Bootparms: "ignore_loglevel"
```

- To configure an NSS LINUX1 as the re-IPL device:

```
# chreipl nss LINUX1
```
chshut - Control the system shutdown actions

Use the **chshut** command to change the shutdown actions for specific shutdown triggers.

The shutdown triggers are:

- `halt`
- `poff`
- `reboot`

The shutdown triggers `restart` and `panic` are handled by the `dumpconf` service script, see *Using the Dump Tools on Red Hat Enterprise Linux 7*, SC34-2711 for details.

Linux on Z performs shutdown actions according to sysfs attribute settings within the `/sys/firmware` directory structure. The **chshut** command sets a shutdown action for a shutdown trigger by changing the corresponding sysfs attribute setting. For more information about the sysfs attributes and the shutdown actions, see Chapter 8, “Shutdown actions,” on page 109.

**chshut syntax**

```
chshut [halt|poff|reboot|ipl|reipl|stop|vmcmd "<cp_command>"]
```

Where:

- **halt**
  - sets an action for the `halt` shutdown trigger.
  
  Red Hat Enterprise Linux by default maps `halt` to `poff`. You can undo this mapping by editing the file `/etc/sysconfig/shutdown` and replacing `HALT="auto"` with `HALT="halt"`.

- **poff**
  - sets an action for the `poff` shutdown trigger.

- **reboot**
  - sets an action for the `reboot` shutdown trigger.

- **ipl**
  - sets IPL as the action to be taken.

- **reipl**
  - sets re-IPL as the action to be taken.

- **stop**
  - sets “stop” as the action to be taken.

- **vmcmd "<cp_command>"**
  - sets the action to be taken to issuing a z/VM CP command. The command must be specified in uppercase characters and enclosed in quotation marks. To issue multiple commands, repeat the `vmcmd` attribute with each command.
-h or --help
    displays help information for the command. To view the man page, enter man
    chshut.

-v or --version
    displays version information.

Examples

These examples illustrate common uses for chshut.

- To make the system start again after a power off:
  ```
  # chshut poff ipl
  ```

- To log off the z/VM guest virtual machine if the Linux poweroff command was
  run successfully:
  ```
  # chshut poff vmcmd LOGOFF
  ```

- To send a message to z/VM user ID OPERATOR and automatically log off the
  z/VM guest virtual machine if the Linux poweroff command is run:
  ```
  # chshut poff vmcmd "MSG OPERATOR Going down" vmcmd "LOGOFF"
  ```
chzcrypt - Modify the cryptographic configuration

Use the `chzcrypt` command to configure cryptographic adapters that are managed by zcrypt and modify zcrypt's AP bus attributes.

To display the attributes, use "lszcrypt - Display cryptographic devices" on page 591.

chzcrypt syntax

```
```

Where:

- **-e** or **--enable**
  sets the given cryptographic devices and AP queues online.

- **-d** or **--disable**
  sets the given cryptographic devices and AP queues offline.

- **-a** or **--all**
  sets all available cryptographic adapters online or offline.

- **<device ID>**
  specifies a cryptographic adapter that is to be set online or offline. A cryptographic adapter can be specified either in decimal notation or hexadecimal notation with a '0x' prefix.

- **-p** or **--poll-thread-enable**
  enables zcrypt's poll thread.

- **-n** or **--poll-thread-disable**
  disables zcrypt's poll thread.

- **-c <timeout>** or **--config-time <timeout>**
  sets configuration timer for rescanning the AP bus to <timeout> seconds.

- **-t <time>** or **--poll-timeout=<time>**
  sets the high-resolution polling timer to <time> nanoseconds. To display the value, use `lszcrypt -b`.

- **-q <domain>** or **--default-domain <domain>**
  changes the default domain. Specify the domain as either a hexadecimal or decimal value.

**Important**: Be sure to enter an existing domain. The Trusted Key Entry (TKE) workstation does not find the cryptographic adapters if a non-existing domain is entered here. All CCA applications use the default domain, and do not work correctly if the specified domain does not exist.
chzcrypt

-V or --verbose
displays verbose messages.

-h or --help
displays short information about command usage.

-v or --version
displays version information.

Examples

These examples illustrate common uses for chzcrypt.

- To set the cryptographic devices with the adapter IDs 0, 1, 4, 5, and 12 and their
  associated AP queues online (in decimal notation):

  chzcrypt -e 0 1 4 5 12

  Or, in hexadecimal notation:

  chzcrypt -e 0x00 0x01 0x04 0x05 0x0C

- To set the cryptographic adapters 0, 1, 4, 5, and 12 online (in decimal notation):

  chzcrypt -e 0 1 4 5 12

- To set all available cryptographic adapters including all AP queues, offline:

  chzcrypt -d -a

- To set the AP queue defined by adapter ID 00 and domain 77 (0x4d) offline:

  chzcrypt -d 00.004d

- To set the configuration timer for rescanning the AP bus to 60 seconds and
disable zcrypt's poll thread:

  chzcrypt -c 60
  chzcrypt -n

- To change the default domain to 77 (0x4d):

  chzcrypt -q 0x4d

  or

  chzcrypt -q 77
Use the cmsfs-fuse command to mount the enhanced disk format (EDF) file system on a z/VM minidisk.

In Linux, the minidisk is represented as a DASD and the file system is mounted as a cmsfs-fuse file system. The cmsfs-fuse file system translates the record-based file system on the minidisk into Linux semantics.

Through the cmsfs-fuse file system, the files on the minidisk become available to applications on Linux. Applications can read from and write to files on minidisks. Optionally, the cmsfs-fuse file system converts text files between EBCDIC on the minidisk and ASCII within Linux.

**Attention:** You can inadvertently damage files and lose data when directly writing to files within the cmsfs-fuse file system. To avoid problems when you write, multiple restrictions must be observed, especially regarding linefeeds (see restrictions for write).

**Tip:** If you are unsure about how to safely write to a file on the cmsfs-fuse file system, copy the file to a location outside the cmsfs-fuse file system, edit the file, and then copy it back to its original location.

Use `fusermount` to unmount file systems that you mounted with `cmsfs-fuse`. See the `fusermount` man page for details.

**Before you begin:**
- The fuse module must be loaded, for example, with `modprobe fuse`.
- The FUSE library must be installed on your system. Install the fuse-libs RPM delivered with Red Hat Enterprise Linux 7.5.
- The DASD must be online.
- Depending whether you intend to read, write, or both, you must have the appropriate permissions for the device node.

### cmsfs-fuse syntax

```
$ cmsfs-fuse [options] <node> <mount-point> [mount-options] [fuse-options]
```

Where:

- `-a` or `--ascii`
  
  treats all files on the minidisk as text files and converts them from EBCDIC to ASCII.

- `-t` or `--filetype`
  
  treats files with extensions as listed in the `cmsfs-fuse` configuration file as text files and converts them from EBCDIC to ASCII.
The `cmsfs-fuse` command uses `~/.cmsfs-fuse/filetypes.conf` as the configuration file. If this file within the user's home directory does not exist, the default configuration file `/etc/cmsfs-fuse/filetypes.conf` is used.

The `filetypes.conf` file lists one file type per line. Lines that start with a number sign (#) followed by a space are treated as comments and are ignored.

`--from <code-page>`

specifies the encoding of the files on the z/VM minidisk. If this option is not specified, code page CP1047 is used. Enter `iconv --list` to display a list of all available code pages.

`--to <code-page>`

specifies the encoding to which the files on the z/VM minidisk are converted in Linux. If this option is not specified, code page ISO-8859-1 is used. Enter `iconv --list` to display a list of all available code pages.

`<mount-options>`

options as available for the `mount` command. See the `mount` man page for details.

`<fuse-options>`

options for FUSE. The following options are supported by the `cmsfs-fuse` command. To use an option, it must also be supported by the version of FUSE that you have.

- `-d` or `--debug`
  enables debug output (implies `-f`).

- `-f` or `--foreground`
  runs the command as a foreground operation.

- `-o allow_other`
  allows access to other users.

- `-o allow_root`
  allows access to root.

- `-o nonempty`
  allows mounts over files and non-empty directories.

- `-o default_permissions`
  enables permission checking by the kernel.

- `-o max_read=<n>`
  sets maximum size of read requests.

- `-o kernel_cache`
  caches files in the kernel.

- `-o [no]auto_cache`
  enables or disables caching based on modification times.

- `-o umask=<mask>`
  sets file permissions (octal).

- `-o uid=<n>`
  sets the file owner.

- `-o gid=<n>`
  sets the file group.

- `-o max_write=<n>`
  sets the maximum size of write requests.
-o max_readahead=<n>
  sets the maximum readahead value.

-o async_read
  performs reads asynchronously (default).

-o sync_read
  performs reads synchronously.

-o big_writes
  enables write operations with more than 4 KB.

<node>
  the device node for the DASD that represents the minidisk in Linux.

<mount-point>
  the mount point in the Linux file system where you want to mount the CMS file system.

-h or --help
  displays help information for the command. To view the man page, enter man cmsfs-fuse.

-v or --version
  displays version information for the command.

You can use the following extended attributes to handle the CMS characteristics of a file:

user.record_format
  specifies the format of the file. The format is F for fixed record length files and V for variable record length files. This attribute can be set only for empty files. The default file format for new files is V.

user.record_lrecl
  specifies the record length of the file. This attribute can be set only for an empty fixed record length file. A valid record length is an integer in the range 1-65535.

user.file_mode
  specifies the CMS file mode of the file. The file mode consists of a mode letter from A-Z and mode number in the range 0 - 6. The default file mode for new files is A1.

You can use the following system calls to work with extended attributes:

listxattr
  to list the current values of all extended attributes.

getxattr
  to read the current value of a particular extended attribute.

setxattr
  to set a particular extended attribute.

You can use these system calls through the getattr and setfattr commands, available from the attr RPM. For more information, see the man pages of these commands and of the listxattr, getxattr, and setxattr system calls.
Restrictions

When you work with files in the cmsfs-fuse file system, restrictions apply for the following system calls:

write  Be aware of the following restrictions when you write to a file on the cmsfs-fuse file system:

Write location
Writing is supported only at the end of a file.

Padding
For fixed-length record files, the last record is padded to make up a full record length. The padding character is zero in binary mode and the space character in ASCII mode.

Sparse files
Sparse files are not supported. To prevent the cp tool from writing in sparse mode specify -sparse=never.

Records and linefeeds with ASCII conversion (-a and -t)
In the ASCII representation of an EBCDIC file, a linefeed character determines the end of a record. Follow these rules about linefeed characters requirements when you write to EBCDIC files in ASCII mode:

For fixed record length files
Use linefeed characters to separate character strings of the fixed record length.

For variable record length files
Use linefeed characters to separate character strings. The character strings must not exceed the maximum record length.

The CMS file system does not support empty records. cmsfs-fuse adds a space to records that consist of a linefeed character only.

rename and creat
Uppercase file names are enforced.

truncate
Only shrinking of a file is supported. For fixed-length record files, the new file size must be a multiple of the record length.

Examples

- To mount the CMS file system on the minidisk represented by the file node /dev/dasde at /mnt:

  # cmsfs-fuse /dev/dasde /mnt

- To mount the CMS file system on the minidisk represented by the file node /dev/dasde at /mnt and enable EBCDIC to ASCII conversion for text files with extensions as specified in ~/cmsfs-fuse/filetypes.conf or /etc/cmsfs-fuse/filetypes.conf if the former does not exist:

  # cmsfs-fuse -t /dev/dasde /mnt

- To mount the CMS file system on the minidisk represented by the file node /dev/dasde at /mnt and allow root to access the mounted file system:
To unmount the CMS file system that was mounted at /mnt:

```
# fusermount -u /mnt
```

To show the record format of a file, PROFILE.EXEC, on a z/VM minidisk that is mounted on /mnt:

```
# getfattr -n user.record_format /mnt/PROFILE.EXEC
```

To set record length 80 for an empty fixed record format file, PROFILE.EXEC, on a z/VM minidisk that is mounted on /mnt:

```
# setfattr -n user.record_lrecl -v 80 /mnt/PROFILE.EXEC
```
**cpuplugd**

**cpuplugd - Control CPUs and memory**

Use the `cpuplugd` command and a set of rules in a configuration file to dynamically enable or disable CPUs. For Linux on z/VM, you can also dynamically add or remove memory.

Rules that are tailored to a particular system environment and the associated workload can increase performance. The rules can include various system load variables.

You can start cpuplugd from the command line in two ways:

- With the service utility
- From a command line

**Note:** Do not run multiple instances of cpuplugd simultaneously.

**cpuplugd service utility syntax**

If you run the `cpuplugd` daemon through the service utility, you configure the daemon through specifications in the `/etc/sysconfig/cpuplugd` configuration file.

```
$ service cpuplugd start
$ service cpuplugd stop
$ service cpuplugd status
$ service cpuplugd restart
```

Where:

**start**

starts the cpuplugd daemon with the configuration in `/etc/sysconfig/cpuplugd`. Do not run multiple instances of cpuplugd simultaneously. Check the cpuplugd status before starting a new instance.

**stop**

stops the cpuplugd daemon.

**status**

shows current status of cpuplugd.

**restart**

stops and restarts the cpuplugd daemon. Useful to re-read the configuration file when it was changed.

**Examples**

- To stop a running instance of cpuplugd:
  ```
  # service cpuplugd stop
  ```

- To display the status:
  ```
  # service cpuplugd status
  ...
  Active: active (running) ...
  ```
cpuplugd command-line syntax

You can start cpuplugd through a command interface.

**Before you begin:** Do not run multiple instances of cpuplugd simultaneously. Check the cpuplugd status through the service utility before you issue the cpuplugd command (see “cpuplugd service utility syntax” on page 510).

**cpuplugd syntax**

```
cpuplugd [-f] [-V] -c <config file>
```

Where:

- `-c <config file> or --config <config file>` specifies the path to the configuration file with the rules (see “Configuration file structure”).

After you install cpuplugd for the first time, you can find a sample configuration file at `/etc/sysconfig/cpuplugd`. If you are upgrading from a prior version of cpuplugd, see “Migrating old configuration files” on page 512.

- `-f or --foreground` runs in foreground.

- `-V or --verbose` displays verbose messages.

- `-h or --help` displays help information for the command. To view the command man page, enter `man cpuplugd`. To view the man page for the configuration file, enter `man cpuplugd.conf`

- `-v or --version` displays version information.

**Examples**

- To start cpuplugd in daemon mode with a configuration file `/etc/sysconfig/cpuplugd`:

  ```
  # cpuplugd -c /etc/sysconfig/cpuplugd
  ```

- To run cpuplugd in the foreground with verbose messages and with a configuration file `/etc/sysconfig/cpuplugd`:

  ```
  # cpuplugd -V -f -c /etc/sysconfig/cpuplugd
  ```

**Configuration file structure**

The cpuplugd configuration file can specify rules for controlling the number of active CPUs and for controlling the amount of memory.

The configuration file contains:

- `<variable>="<value>"` pairs
These pairs must be specified within one line. The maximum valid line length is 2048 characters. The values can be decimal numbers or algebraic or Boolean expressions.

- Comments
  Any part of a line that follows a number sign (#) is treated as a comment. There can be full comment lines with the number sign at the beginning of the line or comments can begin in mid-line.

- Empty lines

**Attention:** These configuration file samples illustrate the syntax of the configuration file. Do not use the sample rules on production systems. Useful rules differ considerably, depending on the workload, resources, and requirements of the system for which they are designed.

**Migrating old configuration files**
With Red Hat Enterprise Linux 6.2, an enhanced version of **cpuplugd** was introduced.

This enhanced version includes extensions to the configuration file and a new sample configuration file, `/etc/sysconfig/cpuplugd`.

If a configuration file from a prior version of cpuplugd already exists at `/etc/sysconfig/cpuplugd`, this file is not replaced but complemented with new variables. The new sample configuration file is then copied to `/var/adm/fillup-templates/sysconfig.cpuplugd`.

The new sample file contains comments that describe the enhanced file layout. View the file to see this information. Consider merging the existing configuration file with a copy of the new sample file to obtain a configuration file with the existing rules, the new variables, and the new descriptions.

**Basic configuration file for CPU control**
A configuration file for dynamically enabling or disabling CPUs has several required specifications.

The configuration file sample of [Figure 78](#) has been reduced to the specifications that are required for dynamically enabling or disabling CPUs.

```
UPDATE="10"
CPU_MIN="2"
CPU_MAX="10"

HOTPLUG = "idle < 10.0"
HOTUNPLUG = "idle > 100"
```

*Figure 78. Simplified configuration file with CPU hotplug rules*

In the configuration file:

**UPDATE**

specifies the time interval, in seconds, at which cpuplugd evaluates the rules and, if a rule is met, enables or disables CPUs. This variable is also required for controlling memory (see “Basic configuration file for memory control” on page 513).

In the example, the rules are evaluated every 10 seconds.
**CPU_MIN**

specifies the minimum number of CPUs. Even if the rule for disabling CPUs is met, cpuplugd does not reduce the number of CPUs to less than this number.

In the example, the number of CPUs cannot become less than 2.

**CPU_MAX**

specifies the maximum number of CPUs. Even if the rule for enabling CPUs is met, cpuplugd does not increase the number of CPUs to more than this number. If 0 is specified, the maximum number of CPUs is the number of CPUs available on the system.

In the example, the number of CPUs cannot become more than 10.

**HOTPLUG**

specifies the rule for dynamically enabling CPUs. The rule resolves to a boolean true or false. Each time this rule is true, cpuplugd enables one CPU, unless the number of CPUs has already reached the maximum specified with CPU_MAX.

Setting HOTPLUG to 0 disables dynamically adding CPUs.

In the example, a CPU is enabled when the idle times of all active CPUs sum up to less than 10.0%. See “Keywords for CPU hotplug rules” on page 515 for information about available keywords.

**HOTUNPLUG**

specifies the rule for dynamically disabling CPUs. The rule resolves to a boolean true or false. Each time this rule is true, cpuplugd disables one CPU, unless the number of CPUs has already reached the minimum specified with CPU_MIN.

Setting HOTUNPLUG to 0 disables dynamically removing CPUs.

In the example, a CPU is disabled when the idle times of all active CPUs sum up to more than 100%. See “Keywords for CPU hotplug rules” on page 515 for information about available keywords.

If one of these variables is set more than once, only the last occurrence is used. These variables are not case sensitive.

If both the HOTPLUG and HOTUNPLUG rule are met simultaneously, HOTUNPLUG is ignored.

**Basic configuration file for memory control**

For Linux on z/VM, you can also use cpuplugd to dynamically add or take away memory. There are several required specifications for memory control.

The configuration file sample of Figure 79 on page 514 was reduced to the specifications that are required for dynamic memory control.
In the configuration file:

**UPDATE**

specifies the time interval, in seconds, at which cpuplugd evaluates the rules and, if a rule is met, adds or removes memory. This variable is also required for controlling CPUs (see “Basic configuration file for CPU control” on page 512).

In the example, the rules are evaluated every 10 seconds.

**CMM_MIN**

specifies the minimum amount of memory, in 4 KB pages, that Linux surrenders to the CMM static page pool (see “Cooperative memory management background” on page 373). Even if the MEMPLUG rule for taking memory from the CMM static page pool and adding it to Linux is met, cpuplugd does not decrease this amount.

In the example, the amount of memory that is surrendered to the static page pool can be reduced to 0.

**CMM_MAX**

specifies the maximum amount of memory, in 4 KB pages, that Linux surrenders to the CMM static page pool (see “Cooperative memory management background” on page 373). Even if the MEMUNPLUG rule for removing memory from Linux and adding it to the CMM static page pool is met, cpuplugd does not increase this amount.

In the example, the amount of memory that is surrendered to the static page pool cannot become more than 131072 pages of 4 KB (512 MB).

**CMM_INC**

specifies the amount of memory, in 4 KB pages, that is removed from Linux when the MEMUNPLUG rule is met. Removing memory from Linux increases the amount that is surrendered to the CMM static page pool.

In the example, the amount of memory that is removed from Linux is 10240 pages of 4 KB (40 MB) at a time.

**CMM_DEC**

Optional: specifies the amount of memory, in 4 KB pages, that is added to Linux when the MEMPLUG rule is met. Adding memory to Linux decreases the amount that is surrendered to the CMM static page pool.

If this variable is omitted, the amount of memory that is specified for CMM_INC is used.

In the example, CMM_DEC is omitted and the amount of memory added to Linux is 10240 pages of 4 KB (40 MB) at a time, as specified with CMM_INC.

**MEMPLUG**

specifies the rule for dynamically adding memory to Linux. The rule resolves to a boolean true or false. Each time this rule is true, cpuplugd adds the

```plaintext
UPDATE="10"
CMM_MIN="0"
CMM_MAX="131072"  # 512 MB
CMM_INC="10240"    # 40 MB
MEMPLUG = "swaprate > 250"
MEMUNPLUG = "swaprate < 10"
```

Figure 79. Simplified configuration file with memory hotplug rules
number of pages that are specified by CMM_DEC, unless the CMM static page pool already reached the minimum that is specified with CMM_MIN.

Setting MEMPLUG to 0 disables dynamically adding memory to Linux.

In the example, memory is added to Linux if there are more than 250 swap operations per second. See “Keywords for memory hotplug rules” on page 516 for information about available keywords.

MEMUNPLUG specifies the rule for dynamically removing memory from Linux. The rule resolves to a boolean true or false. Each time this rule is true, cpuplugd removes the number of pages that are specified by CMM_INC, unless the CMM static page pool already reached the maximum that is specified with CMM_MAX.

Setting MEMUNPLUG to 0 disables dynamically removing memory from Linux.

In the example, memory is removed from Linux when there are less than 10 swap operations per second. See “Keywords for memory hotplug rules” on page 516 for information about available keywords.

If any of these variables are set more than once, only the last occurrence is used. These variables are not case-sensitive.

If both the MEMPLUG and MEMUNPLUG rule are met simultaneously, MEMUNPLUG is ignored.

CMM_DEC and CMM_INC can be set to a decimal number or to a mathematical expression that uses the same algebraic operators and variables as the MEMPLUG and MEMUNPLUG hotplug rules (see “Keywords for memory hotplug rules” on page 516 and “Writing more complex rules” on page 517).

Predefined keywords

There is a set of predefined keywords that you can use for CPU hotplug rules and a set of keywords that you can use for memory hotplug rules. All predefined keywords are case sensitive.

Keywords for CPU hotplug rules:

There are predefined keywords for use in the CPU hotplug rules, HOTPLUG and HOTUNPLUG.

loadavg
   is the current load average.

onumcpus
   is the current number of online CPUs.

runnable_proc
   is the current number of runnable processes.

user
   is the current CPU user percentage.

nice
   is the current CPU nice percentage.

system
   is the current CPU system percentage.
idle
is the current CPU idle percentage.

iowait
is the current CPU iowait percentage.

irq
is the current CPU irq percentage.

softirq
is the current CPU softirq percentage.

steal
is the current CPU steal percentage.

guest
is the current CPU guest percentage.

guest_nice
is the current CPU guest_nice percentage.

cpustat.<name>
is data from /proc/stat and /proc/loadavg. In the keyword, <name> can be any of the previously listed keywords, for example, cpustat.idle. See the proc man page for more details about the data that is represented by these keywords.

With this notation, the keywords resolve to raw timer ticks since system start, not to current percentages. For example, idle resolves to the current idle percentage and cpustat.idle resolves to the total timer ticks spent idle. See “Using historical data” on page 517 about how to obtain average and percentage values.

loadavg, onumcpus, and runnable_proc are not percentages and resolve to the same values as cpustat.loadavg, cpustat.onumcpus, and cpustat.runnable_proc.

cpustat.total_ticks
is the total number of timer ticks since system start.

time
is the UNIX epoch time in the format “seconds.milliseconds”.

Percentage values are accumulated for all online CPUs. Hence, the values for the percentages range from 0 to 100 × (number of online CPUs). To get the average percentage per CPU device, divide the accumulated value by the number of CPUs. For example, idle / onumcpus yields the average idle percentage per CPU.

Keywords for memory hotplug rules:

There are predefined keywords for use in the memory hotplug rules, MEMPLUG and MEMUNPLUG.

apcr
is the number of page cache operations, pgpin + pgpout, from /proc/vmstat in 512-byte blocks per second.

freemem
is the amount of free memory in MB.

swaprate
is the number of swap operations, pswpin + pswpout, from /proc/vmstat in 4 KB pages per second.
meminfo.<name>
is the value for the symbol <name> as shown in the output of
cat /proc/meminfo. The values are plain numbers but refer to the same units
as those used in /proc/meminfo.

vmstat.<name>
is the value for the symbol <name> as shown in the output of
cat /proc/vmstat.

Using historical data:

Historical data is available for the keyword time and the sets of keywords
cpustat.<name>, meminfo.<name>, and vmstat.<name>.

See “Keywords for CPU hotplug rules” on page 515 and “Keywords for memory
hotplug rules” on page 516 for details about these keywords.

Use the suffixes [\<\n\>] to retrieve the data of \<\n\> intervals in the past, where \<\n\>
can be in the range 0 - 100.

Examples
cpustat.idle
  yields the current value for the counted idle ticks.

cpustat.idle[1]
yields the idle ticks as counted one interval ago.

cpustat.idle[5]
yields the idle ticks as counted five intervals ago.

cpustat.idle - cpustat.idle[5]
yields the idle ticks during the past five intervals.

time - time[1]
yields the length of an update interval in seconds.

cpustat.total_ticks - cpustat.total_ticks[5]
yields the total number of ticks during the past five intervals.

(cpustat.idle - cpustat.idle[5]) / (cpustat.total_ticks - cpustat.total_ticks[5])
yields the average ratio of idle ticks to total ticks during the past five
intervals.

Multiplying this ratio with 100 yields the percentage of idle ticks during
the last five intervals.

Multiplying this ratio with 100 * onumcpu yields the accumulated
percentage of idle ticks for all processors during the last five intervals.

Writing more complex rules

In addition to numbers and keywords, you can use mathematical and Boolean
operators, and you can use user-defined variables to specify rules.

- The keywords of “Predefined keywords” on page 515
- Decimal numbers
- The mathematical operators
  + addition
  - subtraction
  * multiplication
  / division
**cpuplugd**

- less than
- greater than
- Parentheses ( and ) to group mathematical expressions
- The Boolean operators
  & and
  | or
  ! not
- User-defined variables
  You can specify complex calculations as user-defined variables, which can then be used in expressions. User-defined variables are case-sensitive and must not match a pre-defined variable or keyword. In the configuration file, definitions for user-defined variables must precede their use in expressions.
  Variable names consist of alphanumeric characters and the underscore (_) character. An individual variable name must not exceed 128 characters. All user-defined variable names and values, in total, must not exceed 4096 characters.

**Examples**
- HOTPLUG = "loadavg > onumcpus + 0.75"
- HOTPLUG = "(loadavg > onumcpus + 0.75) & (idle < 10.0)"
- my_idle_rate = "(cpustat.idle - cpustat.idle[5]) / (cpustat.total_ticks - cpustat.total_ticks[5])"
  my_idle_percent_total = "my_idle_rate * 100 * onumcpus"
- HOTPLUG = "(loadavg > onumcpus + 0.75) & (my_idle_percent_total < 10.0)"

**Sample configuration file**
A typical configuration file includes multiple user-defined variables and values from procfs, for example, to calculate the page scan rate or the cache size.
Attention: The sample file of Figure 80 illustrates the syntax of the configuration file. Useful rules might differ considerably, depending on the workload, resources, and requirements of the system for which they are designed.

After you install cpuplugd with the s390utils RPM, a commented sample configuration file is available at /etc/sysconfig/cpuplugd. This file is used by the cpuplugd service.
dasdfmt

dasdfmt - Format a DASD

Use the dasdfmt command to low-level format ECKD-type direct access storage devices (DASD).

dasdfmt uses an ioctl call to the DASD driver to format tracks. A block size (hard sector size) can be specified. The formatting process can take quite a long time (hours for large DASD).

Tips:
- For DASDs that have previously been formatted with the cd1 or ld1 disk layout, use the dasdfmt quick format mode.
- Use the -p option to monitor the progress.

CAUTION:
As on any platform, formatting irreversibly destroys data on the target disk. Be sure not to format a disk with vital data unintentionally.

dasdfmt syntax

Notes:
1. You are prompted for the block size if you omit it for an option that requires this specification.
2. If neither the -l option nor the -k option are specified, a VOLSER is generated from the device number through which the volume is accessed.

Where:

-r <cylinders> or --requestsdae=<cylinders>
specifies the number of cylinders to be processed in one formatting step. The value must be an integer in the range 1 - 255. The default is 10 cylinders. This parameter exploits any available PAV devices. Ideally, the number of cylinders matches the number of associated devices, counting the base device and all alias devices.

-b <block_size> or --blocksize=<block_size>
specifies one of the following block sizes in bytes: 512, 1024, 2048, or 4096.

For the quick and expand modes and for the --check option, you can omit the block size. Otherwise, you are prompted if you do not specify a value for the block size. You can then press Enter to accept 4096 or specify a different value.
Tip: Set `<block_size>` as large as possible (ideally 4096); the net capacity of an ECKD DASD decreases for smaller block sizes. For example, a DASD formatted with a block size of 512 byte has only half of the net capacity of the same DASD formatted with a block size of 4096 byte.

`<node>` specifies the device node of the device to be formatted, for example, `/dev/dasdzzz`. See “DASD naming scheme” on page 127 for more details about device nodes.

`-d <disklayout> or --disk_layout=<disklayout>` formats the device with the compatible disk layout (cdl) or the Linux disk layout (ldl). If the parameter is not specified, the default (cdl) is used.

`-L or --no_label` valid for `-d ldl` only, where it suppresses the default LNX1 label.

`-l <volser> or --label=<volser>` specifies the volume serial number (see `VOLSER`) to be written to the disk. If the VOLSER contains special characters, it must be enclosed in single quotation marks. In addition, any ‘$’ character in the VOLSER must be preceded by a backslash (`\`).

`-k or --keep_volser` keeps the volume serial number when writing the volume label (see `VOLSER`). Keeping the volume serial number is useful, for example, if the volume serial number was written with a z/VM tool and should not be overwritten.

`-M or --mode=<mode>` specifies the mode to be used for formatting the device. Valid modes are:

- **full** Format the entire disk with the specified block size. This is the default mode.

- **quick** formats the first two tracks and writes label and partition information. Only use this option if you are sure that the target DASD already contains a regular format with the specified block size.

- **expand** format all unformatted tracks at the end of the target DASD. This mode assumes that tracks at the beginning of the DASD volume have already been correctly formatted, while a consecutive set of tracks at the end are unformatted. You can use this mode to make added space available for Linux use after dynamically increasing the size of a DASD volume.

For the **quick** and **expand** modes, omit the block size specification (-b option) to use the existing block size. If you specify a block size, `dasdfmt` checks that the specification matches the existing block size before formatting.

`-p or --progressbar` displays a progress bar. Do not use this option if you are using a line-mode terminal console driver. For example, if you are using a 3215 terminal device driver or a line-mode hardware console device driver.

`-P or --percentage` displays one line for each formatted cylinder. The line shows the number of the cylinder and percentage of formatting process. Intended for use by higher level interfaces.
Dasdfmt

-m <hashstep> or --hashmarks=<hashstep>
    displays a number sign (#) after every <hashstep> cylinders are formatted.
    <hashstep> must be in the range 1 - 1000. The default is 10.

    The -m option is useful where the console device driver is not suitable for the
    progress bar (-p option).

--check
    performs a complete format check on a DASD volume.

    Omit the block size specification (-b option) to check for a consistent format
    for any valid block size. Specify a block size to confirm that the DASD has
    been formatted consistently with that particular block size.

-y starts formatting immediately without prompting for confirmation.

-F or --force
    formats the device without checking whether it is mounted.

-v displays extra information messages (verbose).

-t or --test
    runs the command in test mode. Analyzes parameters and displays what
    would happen, but does not modify the disk.

-C or --check_host_count
    checks the host-access open count to ensure that the device is not online to
    another operating system instance. Use this option to ensure that the operation
    is safe, and cancel it if other operating system instances are accessing the
    volume.

-- norecordzero
    prevents a format write of record zero. This option is intended for experts:
    Subsystems in DASD drivers are by default granted permission to modify or
    add a standard record zero to each track when needed. Before you revoke the
    permission with this option, you must ensure that the device contains standard
    record zeros on all tracks.

-V or --version
    displays the version number of dasdfmt and exits.

-h or --help
    displays an overview of the syntax. Any other parameters are ignored. To view
    the man page, enter man dasdfmt.

Examples

- To format a 100 cylinder z/VM minidisk with the standard Linux disk layout
  and a 4 KB blocksize with device node /dev/dasdc:
# dasdfmt -b 4096 -d ld1 -p /dev/dasdc
Drive Geometry: 100 Cylinders x 15 Heads = 1500 Tracks

I am going to format the device /dev/dasdc in the following way:
  Device number of device : 0x192
  Labelling device : yes
  Disk label : LNX1
  Disk identifier : 0X0192
  Extent start (trk no) : 0
  Extent end (trk no) : 1499
  Compatible Disk Layout : no
  Blocksize : 4096
  Mode 

---> ATTENTION! <<---
All data of that device will be lost.
Type yes to continue, no will leave the disk untouched: yes
Formatting the device. This may take a while (get yourself a coffee).

cyl 100 of 100 ###########################|100% [1s]

Finished formatting the device.
Rereading the partition table... ok
#

- To format the same disk with the compatible disk layout (accepting the default value of the -d option).

# dasdfmt -b 4096 -p /dev/dasdc
Drive Geometry: 100 Cylinders x 15 Heads = 1500 Tracks

I am going to format the device /dev/dasdc in the following way:
  Device number of device : 0x192
  Labelling device : yes
  Disk label : VOL1
  Disk identifier : 0X0192
  Extent start (trk no) : 0
  Extent end (trk no) : 1499
  Compatible Disk Layout : yes
  Blocksize : 4096

---> ATTENTION! <<---
All data of that device will be lost.
Type yes to continue, no will leave the disk untouched: yes
Formatting the device. This may take a while (get yourself a coffee).

cyl 100 of 100 ###################

Finished formatting the device.
Rereading the partition table... ok
#

- To format using the -P option:

# dasdfmt -P /dev/dasdc
cyl 1 of 500 | 0%
cyl 2 of 500 | 0%
cyl 3 of 500 | 0%
cyl 4 of 500 | 0%
cyl 5 of 500 | 1%
...
cyl 496 of 500 | 99%
cyl 497 of 500 | 99%
cyl 498 of 500 | 99%
cyl 499 of 500 | 99%
cyl 500 of 500 | 100%
To make best use of PAV when formatting a DASD that has one base device and four alias devices, specify five cylinders:

```
# dasdfmt /dev/dasdd -y -b 4096 -d cd1 -r 5
Finished formatting the device.
Rereading the partition table... ok
```

To format a previously formatted DASD in quick format mode.

```
# dasdfmt -b 4096 -p --mode=quick /dev/dasdf
```

To format tracks that have been added at the end of an already formatted DASD.

```
# dasdfmt -b 4096 -p --mode=expand /dev/dasdg
```

To check whether a DASD has been correctly formatted with a block size of 4096 bytes.

```
# dasdfmt -b 4096 -p --check /dev/dasdg
Checking format of the entire disk...
cyl 1113 of 1113 |#################################|100% [19s]
Done. Disk is fine.
```

To ensure that the DASD is not online to an operating system instance in a different LPAR when you start formatting the DASD:

```
# dasdfmt -b 4096 -p -C /dev/dasdh
```

dasdfmt always checks the host-access open count. If the count indicates access by another operating system instance, the response depends on the -C option. With this option, the command is canceled. Otherwise, a warning is displayed before you are prompted to confirm that you want to proceed.
dasdstat - Display DASD performance statistics

Use the dasdstat command to display DASD performance statistics, including statistics about Parallel Access Volume (PAV) and High Performance Ficon.

This command includes and extends the performance statistics that is also available through the tunedasd command.

Before you begin: debugfs must be mounted, see “debugfs” on page xi.

dasdstat syntax

Where:

-e or --enable
starts statistics data collection.

-d or --disable
stops statistics data collection.

-r or --reset
sets the statistics counters to zero.

-l or --long
displays more detailed statistics information, for example, differentiates between read and write requests.

-V or --verbose
displays more verbose command information.

-c <colnum> or --columns <colnum>
formats the command output in a table with the specified number of columns. The default is 16. Each row gets wrapped after the specified number of lines.

-w <width> or --column-width <width>
sets the minimum width, in characters, of a column in the output table.

-i <directory> or --directory <directory>
specifies the directory that contains the statistics. The default is <mountpoint>/dasd, where <mountpoint> is the mount point of debugfs. You
dasdstat

need to specify this parameter if the dasdstat command cannot determine this mount point or if the statistics are copied to another location.

<item>
limits the command to the specified items. For <item> you can specify:
- global for summary statistics for all available DASDs.
- The block device name by which a DASD is known to the DASD device driver.
- The bus ID by which a DASD is known as a CCW device. DASDs that are set up for PAV or HyperPAV have a CCW base device and, at any one time, can have one or more CCW alias devices for the same block device. Alias devices are not permanently associated with the same block device. Statistics that are based on bus ID, therefore, show additional detail for PAV and HyperPAV setups.

If you do not specify any individual item, the command applies to all DASD block devices, CCW devices, and to the summary.

-v or --version
displays the version number of dasdstat, then exits.

-h or --help
displays help information for the command.

Examples
- This command starts data collection for dasda, 0.0.b301, and for a summary of all available DASDs.

```
# dasdstat -e dasda 0.0.b301 0.0.b302 global
```

- This command resets the statistics counters for dasda.

```
# dasdstat -r dasda
```

- This command reads the summary statistics:
For details about the data items, see “Interpreting the data rows” on page 145.
Use the `dasdview` command to display DASD information.

`dasdview` displays:
- The volume label.
- VTOC details (general information, and the DSCBs of format 1, format 3, format 4, format 5, format 7, format 8, and format 9).
- The content of the DASD, by specifying:
  - Starting point
  - Size

You can display these values in hexadecimal, EBCDIC, and ASCII format.
- Whether the data on the DASD is encrypted.
- Whether the disk is a solid-state device.

If you specify a start point and size, you can also display the contents of a disk dump.

For more information about partitioning, see "The IBM label partitioning scheme" on page 122.

**dasdview syntax**

```
dasdview [-b <begin> -s <size> -1 -i -x -j -l -c -t <spec> <node>]
```

Where:
- `-b <begin>` or `--begin=<begin>`
  displays disk content on the console, starting from `<begin>`. The contents of the disk are displayed as hexadecimal numbers, ASCII text, and EBCDIC text. If `<size>` is not specified, `dasdview` takes the default size (128 bytes). You can specify the variable `<begin>` as:
  `<begin>[k|m|b|t|c]`

If the disk is in raw-track access mode, you can specify only track (t) or cylinder (c) entities.

The default for `<begin>` is 0.

`dasdview` displays a disk dump on the console with the DASD driver. The DASD driver might suppress parts of the disk, or add information that is not relevant. This discrepancy might occur, for example, when `dasdview` displays the first two tracks of a disk that was formatted with compatible disk layout option (`-d cdl`). In this situation, the DASD driver pads shorter blocks with
zeros to maintain a constant blocksize. All Linux applications (including dasdview) process according to this rule.

Here are some examples of how this option can be used:

- `b 32` (start printing at Byte 32)
- `b 32k` (start printing at kByte 32)
- `b 32m` (start printing at MByte 32)
- `b 32b` (start printing at block 32)
- `b 32t` (start printing at track 32)
- `b 32c` (start printing at cylinder 32)

- `s <size>` or `--size=<size>`
displays a disk dump on the console, starting at `<begin>`, and continuing for size=<size>. The contents of the dump are displayed as hexadecimal numbers, ASCII text, and EBCDIC text. If a start value, `<begin>`, is not specified, dasdview takes the default. You can specify the variable `<size>` as:

size[k|m|b|t|c]

If the disk is in raw-track access mode, you can specify only track (t) or cylinder (c) entities.

The default for `<size>` is 128 bytes.

Here are some examples of how this option can be used:

- `s 16` (use a 16 Byte size)
- `s 16k` (use a 16 kByte size)
- `s 16m` (use a 16 MByte size)
- `s 16b` (use a 16 block size)
- `s 16t` (use a 16 track size)
- `s 16c` (use a 16 cylinder size)

- `1` displays the disk dump with format 1 (as 16 Bytes per line in hexadecimal, ASCII and EBCDIC). A line number is not displayed. You can use option `-1` only together with `-b` or `-s`.

Option `-1` is the default.

- `2` displays the disk dump with format 2 (as 8 Bytes per line in hexadecimal, ASCII and EBCDIC). A decimal and hexadecimal byte count are also displayed. You can use option `-2` only together with `-b` or `-s`.

- `i` or `--info`
displays basic information such as device node, device bus-ID, device type, or geometry data.

- `x` or `--extended`
displays the information that is obtained by using `-i` option, but also open count, subchannel identifier, and so on.

- `j` or `--volser`
prints volume serial number (volume identifier).

- `l` or `--label`
displays the volume label.

- `c` or `--characteristics`
displays model-dependent device characteristics, for example disk encryption status or whether the disk is a solid-state device.

- `t <spec>` or `--vtoc=<spec>`
displays the VTOC’s table-of-contents, or a single VTOC entry, on the console. The variable `<spec>` can take these values:

info displays overview information about the VTOC, such as a list of the data set names and their sizes.
dasdview

f1 displays the contents of all format 1 data set control blocks (DSCBs).
f3 displays the contents of all (z/OS-specific) format 3 DSCBs.
f4 displays the contents of all format 4 DSCBs.
f5 displays the contents of all format 5 DSCBs.
f7 displays the contents of all format 7 DSCBs.
f8 displays the contents of all format 8 DSCBs.
f9 displays the contents of all format 9 DSCBs.
all displays the contents of all DSCBs.

<node> specifies the device node of the device for which you want to display information, for example, /dev/dasdzzz. See “DASD naming scheme” on page 127 for more details about device nodes.

-h or --help displays short usage text on console. To view the man page, enter man dasdview.

-v or --version displays version number on console, and exit.

Examples

• To display basic information about a DASD:

  # dasdview -i /dev/dasdzzz

  This example displays:

  --- general DASD information -----------------------------------------------
  device node : /dev/dasdzzz
  busid : 0.0.0193
  type : ECKD
  device type : hex 3390 dec 13200

  --- DASD geometry -----------------------------------------------------------
  number of cylinders : hex 64 dec 100
  tracks per cylinder : hex f dec 15
  blocks per track : hex c dec 12
  blocksize : hex 1000 dec 4096

• To display device characteristics:

  # dasdview -c /dev/dasda

  This example displays:

  encrypted disk : no
  solid state device : no

• To include extended information:

  # dasdview -x /dev/dasdzzz

  This example displays:
--- general DASD information ---------------------------------------
device node  : /dev/dasdz
busid       : 0.0.0.0193
type        : ECKD
device type : hex 3390  dec 13200
--- DASD geometry --------------------------------------------------
number of cylinders : hex 64  dec 100
tracks per cylinder : hex f  dec 15
blocks per track   : hex c  dec 12
blocksize          : hex 1000  dec 4096
--- extended DASD information -------------------------------------
real device number : hex 452bc08  dec 72530952
CU type (SenseID)   : hex 3990  dec 14736
CU model (SenseID)  : hex e9  dec 233
device type (SenseID) : hex 3390  dec 13200
device model (SenseID) : hex a  dec 10
open count         : hex 1  dec 1
req_queue_len      : hex 0  dec 0
chang_len          : hex 0  dec 0
status             : hex 5  dec 5
label_block        : hex 40  dec 64
FBA_layout         : hex 0  dec 0
characteristics_size : hex 100  dec 256
confdata_size      : hex 00  dec 0
characteristics    : 3990e933 900a5f80 dff22024 00640000
                    e000e5a2 05940222 13090674 00000000
                    00000000 24241502 dffee001
                    0677080f 007f4a00 1b350000 00000000
configuration_data : dc010100 4040f2f1 f0540400 40c9c2d4
                    f1f3f0f0 f0f0f0f0 f0c6c3f1 f1f3f0f0
                    dc000000 4040f2f1 f0540400 40c9c2d4
                    f1f3f0f0 f0f0f0f0 f0c6c3f1 f1f3f0f0
                    d4020000 4040f2f1 f0f0f0f0 f0c9c2d4
                    f1f3f0f0 f0f0f0f0 f0c6c3f1 f1f3f0f0
                    f1f3f0f0 f0f0f0f0 f0c6c3f1 f1f3f0f0
                    00000000 00000000 00000000 00000000
                    00000000 00000000 00000000 00000000
                    00000000 00000000 00000000 00000000
                    00000000 00000000 00000000 00000000
                    00000000 00000000 00000000 00000000
                    00000000 00000000 00000000 00000000
                    00000000 00000000 00000000 00000000
                    800000a1 0001e00 51400009 0909a188
                    0140c09 7cb7efb7 00000000 00000000

#

• To display volume label information:

  # dasdview -l /dev/dasdz

This displays:
--- volume label -----------------------------------------------

volume label key:
| ascii | 'ÅÖÖñ' | ebcdic | 'VOL1' | hex | e5d6d3f1 |

volume label identifier:
| ascii | 'ðçðñùó' | ebcdic | '0X0193' | hex | f0e7f0f1f9f3 |

security byte:
| hex | 40 |

VTOC pointer:
| hex | 0000000101 |
|      | (cyl 0, trk 1, blk 1) |

reserved:
| ascii | '@@@@@' | ebcdic | '@@@@@' | hex | 4040404040 |

CI size for FBA:
| ascii | '@@@@@' | ebcdic | '@@@@@' | hex | 40404040 |

blocks per CI (FBA):
| ascii | '@@@@@' | ebcdic | '@@@@@' | hex | 40404040 |

labels per CI (FBA):
| ascii | '@@@@@' | ebcdic | '@@@@@' | hex | 40404040 |

reserved:
| ascii | '@@@@@' | ebcdic | '@@@@@' | hex | 40404040 |

owner code for VTOC:
| ascii | '@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@}

To display partition information:

```
# dasdview -t info /dev/dasdzzz
```

This example displays:
The VTOC contains:

- 3 format 1 label(s)
- 1 format 4 label(s)
- 1 format 5 label(s)
- 0 format 7 label(s)
- 0 format 8 label(s)
- 0 format 9 label(s)

Other S/390 and zSeries operating systems would see the following data sets:

<table>
<thead>
<tr>
<th>data set</th>
<th>start</th>
<th>end</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINUX.V0X0193.PART0001.NATIVE</td>
<td>trk</td>
<td>trk</td>
</tr>
<tr>
<td>data set serial number</td>
<td>'OX0193'</td>
<td>2</td>
</tr>
<tr>
<td>system code</td>
<td>'IBM LINUX'</td>
<td>cyl/trk</td>
</tr>
<tr>
<td>creation date</td>
<td>year 2001, day 317</td>
<td>0/2</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>LINUX.V0X0193.PART0002.NATIVE</td>
<td>trk</td>
<td>trk</td>
</tr>
<tr>
<td>data set serial number</td>
<td>'OX0193'</td>
<td>501</td>
</tr>
<tr>
<td>system code</td>
<td>'IBM LINUX'</td>
<td>cyl/trk</td>
</tr>
<tr>
<td>creation date</td>
<td>year 2001, day 317</td>
<td>33/6</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>LINUX.V0X0193.PART0003.NATIVE</td>
<td>trk</td>
<td>trk</td>
</tr>
<tr>
<td>data set serial number</td>
<td>'OX0193'</td>
<td>901</td>
</tr>
<tr>
<td>system code</td>
<td>'IBM LINUX'</td>
<td>cyl/trk</td>
</tr>
<tr>
<td>creation date</td>
<td>year 2001, day 317</td>
<td>60/1</td>
</tr>
</tbody>
</table>

To display VTOC information:

```
# dasdview -t f4 /dev/dasdzzz
```

This example displays:
--- VTOC format 4 label --------------------------------------------------------

**DS4KEYCD**: 040404040404040404040404040404040404040404040404040404040404040404...

**DS4IDFMT**: dec 244, hex f4

**DS4HPCHR**: 000000000105 (cyl 0, trk 1, blk 5)

**DS4DSREC**: dec 7, hex 0007

**DS4HCCHH**: dec 0, hex 00

**DS4NOATK**: dec 0, hex 00

**DS4NOEXT**: dec 1, hex 01

**DS4VTOCI**: dec 0, hex 00

**DS4AMTIM**: hex 0000000000000000

**DS4AMCAT**: hex 000000

**DS4R2TIM**: hex 0000000000000000

**DS4F6PTR**: hex 0000000000

**DS4F6PTR**: hex 0000000000

**DS4VTOCE**: hex 010000000100000001
typeind : dec 1, hex 01
seqno : dec 0, hex 00
llimit : hex 00000001 (cyl 0, trk 1)
ulimit : hex 00000001 (cyl 0, trk 1)

**res2**: hex 00000000000000000000

**DS4EFLVL**: dec 0, hex 00

**DS4EFPTR**: hex 0000000000

**DS4FLVL**: dec 0, hex 00

**DS4VTOCE**: hex 010000000100000001
typeind : dec 1, hex 01
seqno : dec 0, hex 00
llimit : hex 00000001 (cyl 0, trk 1)
ulimit : hex 00000001 (cyl 0, trk 1)

**res3**: hex 00000000000000000000

---

To print the contents of a disk to the console starting at block 2 (volume label):

```bash
# dasdview -b 2b -s 128 /dev/dasdzzz
```

This example displays:

```
+----------------------------------------+------------------+------------------+
<p>| HEXADECIMAL                           | EBCDIC           | ASCII            |</p>
<table>
<thead>
<tr>
<th>01....04 05....08 09....12 13....16</th>
<th>1.............16</th>
<th>1.............16</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5D6D3F1 E5D6D3F1 F9F34000</td>
<td>VOLI VOLI 10X193?</td>
<td>???????????????</td>
</tr>
<tr>
<td>000000101 40404040 40404040 40404040</td>
<td>........................</td>
<td>........................</td>
</tr>
<tr>
<td>40404040 40404040 40404040 40404040</td>
<td>........................</td>
<td>........................</td>
</tr>
<tr>
<td>40404040 40404040 40404040 40404040</td>
<td>........................</td>
<td>........................</td>
</tr>
<tr>
<td>40404040 40404040 40404040 40404040</td>
<td>........................</td>
<td>........................</td>
</tr>
<tr>
<td>40404040 40404040 40404040 40404040</td>
<td>........................</td>
<td>........................</td>
</tr>
<tr>
<td>40404040 88001000 10000000 00808000</td>
<td>........................</td>
<td>........................</td>
</tr>
<tr>
<td>00000000 00000000 00010000 00000200</td>
<td>........................</td>
<td>........................</td>
</tr>
<tr>
<td>21000500 00000000 00000000 00000000</td>
<td>........................</td>
<td>........................</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
</tbody>
</table>
```

To display the contents of a disk on the console starting at block 14 (first FMT1 DSCB) with format 2:

```bash
# dasdview -b 14b -s 128 -2 /dev/dasdzzz
```

This example displays:
To see what is at block 1234 (in this example there is nothing there):

```bash
# dasdview -b 1234b -s 128 /dev/dasdzzz
```

This example displays:

```
+----------------------------------------+----------------+------------------+
|                                      |                |                  |
|                                      |                |                  |
|                                      |                |                  |
```

To try byte 0 instead:

```bash
# dasdview -b 0 -s 64 /dev/dasdzzz
```

This example displays:

```
+----------------------------------------+----------------+------------------+
|                                      |                |                  |
|                                      |                |                  |
|                                      |                |                  |
```

To display the contents of a disk on the console starting at cylinder 2 and printing one track of data:

```bash
Chapter 53. Commands 535
```
This example displays:

<table>
<thead>
<tr>
<th>HEXADECIMAL</th>
<th>EBCDIC</th>
<th>ASCII</th>
</tr>
</thead>
<tbody>
<tr>
<td>01..04 05..08 09..12 13..16</td>
<td>1..16</td>
<td>1..16</td>
</tr>
<tr>
<td>52B7DBEE D6B9530B 0179F420 C86EA9SE</td>
<td>???????????z;</td>
<td>R??????y??n?^</td>
</tr>
<tr>
<td>EF49C03C 513542E7 D8F17D9D 06DC44F7</td>
<td>??(????XQ1'?????</td>
<td>?I?q5B???)??D?</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>92963D5B 0200B0FA 53745C12 C3B45125</td>
<td>ko?$.??????</td>
<td>??=[?n????</td>
</tr>
<tr>
<td>0D6040C2 F9333B1E 7A4C4797 F40FEDAB</td>
<td>?-789????:&lt;p????</td>
<td>??0??????zL?????</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

To display the full record information of the same disk when it in raw-track access mode:

```
# dasdview -b 2c -s 1t /dev/dasdk
```
To display the contents of a disk, which is in raw-access mode, printing one track of data from the start of the disk:
This example displays:

cylinder 0, head 0, record 0
+---------------------------------------------------------------+
  count area:
    hex: 0000000000000008
    cylinder: 0
    head: 0
    record: 0
    key length: 0
    data length: 8
+---------------------------------------------------------------+
  key area:
    HEXADECIMAL
    01....04 05....08 09....12 13....16        1.............16 1.............16
+---------------------------------------------------------------+
  data area:
    HEXADECIMAL
    01....04 05....08 09....12 13....16        1.............16 1.............16
+---------------------------------------------------------------+
    00000000 00000000
+---------------------------------------------------------------+

cylinder 0, head 0, record 1
+---------------------------------------------------------------+
  count area:
    hex: 0000000001040018
    cylinder: 0
    head: 0
    record: 1
    key length: 4
    data length: 24
+---------------------------------------------------------------+
  key area:
    HEXADECIMAL
    01....04 05....08 09....12 13....16        1.............16 1.............16
+---------------------------------------------------------------+
    C9D7D3F1
    IPL1............  ???.????????
+---------------------------------------------------------------+
  data area:
    HEXADECIMAL
    01....04 05....08 09....12 13....16        1.............16 1.............16
+---------------------------------------------------------------+
    000A0000 0000000F 03000000 00000001
    ??.??.??.??    ??.??.??.??
    00000000 00000000
+---------------------------------------------------------------+
...
fdasd – Partition a DASD

Use the **fdasd** command to manage partitions on ECKD-type DASD that were formatted with the compatible disk layout.

See “dasdfmt - Format a DASD” on page 520 for information about formatting a DASD. With **fdasd** you can create, change, and delete partitions, and also change the volume serial number.

**fdasd** checks that the volume has a valid volume label and VTOC. If either is missing or incorrect, **fdasd** re-creates it. See “IBM Z compatible disk layout” on page 123 for details about the volume label and VTOC.

Calling **fdasd** with a node, but without options, enters interactive mode. In interactive mode, you are given a menu through which you can display DASD information, add or remove partitions, or change the volume identifier. Your changes are not written to disk until you type the write option on the menu. You can quit without altering the disk at any time before this.

For more information about partitions, see “The IBM label partitioning scheme” on page 122.

**Before you begin:**
- To partition a SCSI disk, use **fdisk** rather than **fdasd**.
- The disk must be formatted with **dasdfmt**, using the compatible disk layout.

**Attention:** Careless use of **fdasd** can result in loss of data.

**fdasd syntax**

```
--fdasd
```

**Notes:**
1 If neither the -1 option nor the -k option is specified, a VOLSER is generated from the device number through which the volume is accessed.

Where:
- **-h** or **--help**
  displays help on command line arguments.
- **-v** or **--version**
  displays the version of **fdasd**.
- **-s** or **--silent**
  suppresses messages.
-r or --verbose
displays additional messages that are normally suppressed.

-a or --auto
auto-creates one partition using the whole disk in non-interactive mode.

-k or --keep_volser
keeps the volume serial number when writing the volume label (see VOLSER).
Keeping the serial number is useful, for example, if the volume serial number
was written with a z/VM tool and should not be overwritten.

-l <volser> or --label <volser>
specifies the volume serial number (see VOLSER).
A volume serial consists of one through six alphanumeric characters or the
following special characters:

$ # @ %

All other characters are ignored. Avoid using special characters in the volume
serial. Special characters can cause problems when accessing a disk by
VOLSER. If you must use special characters, enclose the VOLSER in single
quotation marks. In addition, any '$' character in the VOLSER must be
preceded by a backslash ('\').

For example, specify:
-l 'a@b$c#'
to get:
A@B$C#

VOLSER is interpreted as an ASCII string and is automatically converted to
uppercase, padded with blanks and finally converted to EBCDIC before it is
written to disk.

Do not use the following reserved volume serials:
• SCRTCH
• PRIVAT
• MIGRAT
• Lnnnnn (L followed by a five-digit number)
These volume serials are used as keywords by other operating systems (z/OS).
Omitting this parameter causes fdasd to prompt for it, if it is needed.

-c <conf_file> or --config <conf_file>
creates several partitions in non-interactive mode, according to specifications in
the plain-text configuration file <conf_file>.

For each partition you want to create, add one line of the following format to
<conf_file>:
[first_track], [last_track], [type]

[first_track] and [last_track] are required and specify the first and last track of
the partition. You can use the keyword first for the first possible track on disk
and, correspondingly, the keyword last for the last possible track on disk.
[type] describes the partition type and is one of:

native
for partitions to be used for Linux file systems.
gpfs
for partitions to be used as part of an Elastic Storage file system setup.

swap
for partitions to be used as swap devices.

raid
for partitions to be used as part of a RAID setup.

lvm
for partitions to be used as part of a logical volume group.

The type specification is optional. If the type is omitted, native is used.

The type describes the intended use of a partition to tools or other operating systems. For example, swap partitions could be skipped by backup programs. How Linux actually uses the partition depends on how the partition is formatted and set up. For example, a partition of type native can still be used in an LVM logical volume or in a RAID configuration.

Example: With the following sample configuration file you can create three partitions:

[first,1000,raid]
[1001,2000,swap]
[2001,last]

-i or --volser
displays the volume serial number and exits.

-p or --table
displays the partition table and exits.

<node>
specifies the device node of the DASD you want to partition, for example, /dev/dasdzzz. See “DASD naming scheme” on page 127 for more details about device nodes.

-C or --check_host_count
checks the host-access open count to ensure that the device is not online to another operating system instance. The operation is canceled if another operating system instance is accessing the device.

fdasd menu
If you call fdasd in the interactive mode (that is, with just a node), a menu is displayed.

<table>
<thead>
<tr>
<th>Command</th>
<th>action</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>print this menu</td>
</tr>
<tr>
<td>p</td>
<td>print the partition table</td>
</tr>
<tr>
<td>n</td>
<td>add a new partition</td>
</tr>
<tr>
<td>d</td>
<td>delete a partition</td>
</tr>
<tr>
<td>v</td>
<td>change volume serial</td>
</tr>
<tr>
<td>t</td>
<td>change partition type</td>
</tr>
<tr>
<td>r</td>
<td>re-create VTOC and delete all partitions</td>
</tr>
<tr>
<td>u</td>
<td>re-create VTOC re-using existing partition sizes</td>
</tr>
<tr>
<td>s</td>
<td>show mapping (partition number - data set name)</td>
</tr>
<tr>
<td>q</td>
<td>quit without saving changes</td>
</tr>
<tr>
<td>w</td>
<td>write table to disk and exit</td>
</tr>
</tbody>
</table>

Command (m for help):
Menu commands

Use the `fdasd` menu commands to modify or view information about DASDs.

- **m** re-displays the `fdasd` command menu.
- **p** displays information about the DASD and any partitions on the DASD.

**DASD information:**
- Number of cylinders
- Number of tracks per cylinder
- Number of blocks per track
- Block size
- Volume label
- Volume identifier
- Number of partitions defined

**Partition information:**
- Linux node
- Start track
- End track
- Number of tracks
- Partition ID
- Partition type

There is also information about the free disk space that is not used for a partition.

- **n** adds a partition to the DASD. You are asked to give the start track and the length or end track of the new partition.
- **d** deletes a partition from the DASD. You are asked which partition to delete.
- **v** changes the volume identifier. You are asked to enter a new volume identifier. See `VOLSER` for the format.
- **t** changes the partition type. You are asked to identify the partition to be changed. You are then asked for the new partition type (Linux native or swap). This type is a guideline; the actual use Linux makes of the partition depends on how it is defined with the `mkswap` or `mkfs` tools. The main function of the partition type is to describe the partition to other operating systems. Then, for example, swap partitions can be skipped by backup programs.
- **r** re-creates the VTOC and deletes all partitions.
- **u** re-creates all VTOC labels without removing all partitions. Existing partition sizes are reused. This option is useful to repair damaged labels or migrate partitions that are created with older versions of `fdasd`.
- **s** displays the mapping of partition numbers to data set names. For example:
Example using the menu

This example shows how to use *fdasd* to create two partitions on a z/VM minidisk, change the type of one of the partitions, save the changes, and check the results.

This example shows you how to format a z/VM minidisk with the compatible disk layout. The minidisk has device number 193.

1. Call *fdasd*, specifying the minidisk:

```bash
# fdasd /dev/dasdzzz
```

*fdasd* reads the existing data and displays the menu:

```
reading volume label: VOL1
reading vtoc : ok

Command action
  m print this menu
  p print the partition table
  a add a new partition
  d delete a partition
  v change volume serial
  t change partition type
  r re-create VTDC and delete all partitions
  u re-create VTDC re-using existing partition sizes
  s show mapping (partition number - data set name)
  q quit without saving changes
  w write table to disk and exit

Command (m for help):
```

2. Use the *p* option to verify that no partitions are created yet on this DASD:
3. Define two partitions, one by specifying an end track and the other by specifying a length. (In both cases the default start tracks are used):

   `Command (m for help): m
   First track (1 track = 48 KByte) ([2]-1499): 2
   Using default value 2
   Last track or +size[c|k|M] (2-[1499]): 700
   You have selected track 700`  

   `Command (m for help): n
   First track (1 track = 48 KByte) ([701]-1499): 701
   Using default value 701
   Last track or +size[c|k|M] (701-[1499]): +400
   You have selected track 1100`  

4. Check the results by using the `p` option:

   `Command (m for help): p
   Disk /dev/dasdzzz:
   cylinders ............: 100
   tracks per cylinder ..: 15
   blocks per track ......: 12
   bytes per block ......: 4096
   volume label ..........: VOL1
   volume serial .......: 0X0193
   max partitions .......: 3
   --------------------------------- tracks ------------------------------
   Device    start    end   length    Id      System
   /dev/dasdzzz 2 700 699 1 Linux native
   /dev/dasdzzz 701 1100 400 2 Linux native
   1101 1499 399 unused`  

5. Change the type of a partition:
Command (m for help): t

Disk /dev/dasdzzz:
cylinders ..............: 100
tracks per cylinder ..: 15
blocks per track ......: 12
bytes per block ......: 4096
volume label ..........: VOL1
volume serial ..........: 0X0193
max partitions .......: 3

---------- tracks ----------
Device start end length Id System
/dev/dasdzzz1 2 700 699 1 Linux native
/dev/dasdzzz2 701 1100 400 2 Linux native
1101 1499 399 unused

change partition type
partition id (use 0 to exit):

Enter the ID of the partition you want to change; in this example partition 2:

partition id (use 0 to exit): 2

6. Enter the new partition type; in this example type 2 for swap:
current partition type is: Linux native
1 Linux native
2 Linux swap
new partition type: 2

7. Check the result:

Command (m for help): p

Disk /dev/dasdzzz:
cylinders ..............: 100
tracks per cylinder ..: 15
blocks per track ......: 12
bytes per block ......: 4096
volume label ..........: VOL1
volume serial ..........: 0X0193
max partitions .......: 3

---------- tracks ----------
Device start end length Id System
/dev/dasdzzz1 2 700 699 1 Linux native
/dev/dasdzzz2 701 1100 400 2 Linux swap
1101 1499 399 unused

8. Write the results to disk with the w option:

Command (m for help): w
writing VTOC...
rereading partition table...
#

Example using options
You can partition a DASD by using the -a or -c option without entering the menu mode.
This method is useful for partitioning by using scripts, for example, if you need to partition several hundred DASDs.

With the `-a` option you can create one large partition on a DASD:

```
# fdasd -a /dev/dasdzzz
auto-creating one partition for the whole disk...
writing volume label...
writing VTOC...
rereading partition table...
```

This example creates a partition as follows:

<table>
<thead>
<tr>
<th>Device</th>
<th>start</th>
<th>end</th>
<th>length</th>
<th>Id</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>/dev/dasdzzz1</td>
<td>2</td>
<td>1499</td>
<td>1498</td>
<td>1</td>
<td>Linux native</td>
</tr>
</tbody>
</table>

Using a configuration file that you can create several partitions. For example, the following configuration file, config, creates three partitions:

```
[first,500]
[501,1100]
[1101,last]
```

Submitting the command with the `-c` option creates the partitions:

```
# fdasd -c config /dev/dasdzzz
parsing config file 'config'...
writing volume label...
writing VTOC...
rereading partition table...
```

This example creates partitions as follows:

<table>
<thead>
<tr>
<th>Device</th>
<th>start</th>
<th>end</th>
<th>length</th>
<th>Id</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>/dev/dasdzzz1</td>
<td>2</td>
<td>500</td>
<td>499</td>
<td>1</td>
<td>Linux native</td>
</tr>
<tr>
<td>/dev/dasdzzz2</td>
<td>501</td>
<td>1100</td>
<td>600</td>
<td>2</td>
<td>Linux native</td>
</tr>
<tr>
<td>/dev/dasdzzz3</td>
<td>1101</td>
<td>1499</td>
<td>399</td>
<td>3</td>
<td>Linux native</td>
</tr>
</tbody>
</table>
hmcdrvfs - Mount a FUSE file system for remote access to media in the HMC media drive

Use the **hmcdrvfs** command for read-only access to contents in a DVD, CD, or USB-attached storage in the media drive of an HMC.

**Before you begin:**

- The fuse.hmcdrvfs file system needs access to device node /dev/hmcdrv. This node is created automatically when the hmcdrv kernel module is loaded, see Chapter 27, “HMC media device driver,” on page 355.
- On the HMC, the media must be assigned to the associated system image (use menu Access Removable Media).
- In a z/VM environment, the z/VM guest virtual machine must have at least privilege class B. The media must be assigned to the LPAR where the z/VM hypervisor runs.
- For Linux in LPAR mode, the LPAR activation profile must allow issuing SCLP requests.

With the media assigned to your Linux instance, this command creates a fuse.hmcdrvfs file system with the media content at the specified mount point.

To unmount file systems that you mounted with **hmcdrvfs**, you can use **fusermount**, whether root or non-root user. See the **fusermount** man page for details.

**hmcdrvfs syntax**

```
$hmcdrvfs$ <mount-point> [ <fuse.hmcdrvfs-options> ] [ <mount-options> ] [ <fuse-options> ]
```

Where:

- **-o** or **--opt**
  FUSE or mount command options; for the FUSE options see the following lists, for mount options see the **mount** man page.

**<fuse.hmcdrvfs-options>**

options specific to the fuse.hmcdrvfs file system:

- **-o hmclang=<language>**
  specifies the language setting on the HMC; for valid values, see the **locale** man page.

- **-o hmctz=<time zone>**
  specifies the time zone setting on the HMC; for valid values, see the **tzset** man page.

**<mount-options>**

options as available for the **mount** command. See the **mount** man page for details.
Options for FUSE. The following options are supported by the `cmsfs-fuse` command. To use an option, it must also be supported by the version of FUSE that you have.

- `-d` or `-o debug`
  enables debug output (implies `-f`).

- `-f`
  runs the command as a foreground operation.

- `-s`
  disables multi-threaded operation.

- `-o allow_other`
  allows access to the file system by other users.

- `-o allow_root`
  allows access to the file system by root.

- `-o nonempty`
  allows mounts over files and non-empty directories.

- `-o default_permissions`
  enables permission checking by the kernel.

- `-o fsname=<name>`
  sets the file system name.

- `-o subtype=<type>`
  sets the file system type.

- `-o max_read=<n>`
  sets maximum size of read requests.

- `-o direct_io`
  uses direct I/O.

- `-o kernel_cache`
  caches files in the kernel.

- `-o [no]auto_cache`
  enables or disables caching based on modification times.

- `-o umask=<mask>`
  sets file permissions (octal).

- `-o uid=<n>`
  sets the file owner.

- `-o gid=<n>`
  sets the file group.

- `-o entry_timeout=<secs>`
  sets the cache timeout for names. The default is 1.0 second.

- `-o attr_timeout=<secs>`
  sets the cache timeout for attributes. The default is 1.0 second.

- `-o ac_attr_timeout=<secs>`
  sets the auto cache timeout for attributes. The default is the `attr_timeout` value.

- `-o max_readahead=<n>`
  sets the maximum read ahead value.

- `-o async_read`
  performs reads asynchronously (default).
-o sync_read
  performs reads synchronously.

-o no_remote_lock
  disables remote file locking.

-o intr
  allows requests to be interrupted

-o intr_signal=<num>
  specifies the signal to send on interrupt.

-v or --version
  displays version information for the command.

-h or --help
  displays a short help text, then exits. To view the man page, enter man hmcdrvfs.

The following options for mount policy can be set in the file /etc/ fuse.conf file:

mount_max=<number>
  sets the maximum number of FUSE mounts allowed for non-root users. The default is 1000.

user_allow_other
  allows non-root users to specify the allow_other or allow_root mount options.

Examples

- To mount the contents of the HMC media drive at /mnt/hmc without any special options, use:

  # hmcdrvfs /mnt/hmc

- If the hmcdrv kernel module is not loaded, load it before you issue the hmcdrvfs command:

  # modprobe hmcdrv
  # hmcdrvfs /mnt/hmc

- To translate the UID and GID of files on the HMC media drive to your system users and groups along with overriding the permissions, issue, for example:

  # hmcdrvfs /mnt/hmc -o uid=500 -o gid=1000 -o umask=0337

- To speed up transfer rates to frequently accessed directories, use the cache timeout option:

  # hmcdrvfs /mnt/hmc -o entry_timeout=60

- If the HMC is in a different timezone and is configured for a different language use, for example:

  # hmcdrvfs /mnt/hmc -o hmclang=de_DE -o hmctz=Europe/Berlin

- To also disregard any Daylight Saving Time, specifying hours west of the Prime Meridian (Coordinated Universal Time):

  # hmcdrvfs /mnt/hmc -o hmclang=de_DE -o hmctz="GMT-1"
To unmount the HMC media drive contents mounted on `/mnt/hmc`, issue:

```
# fusermount -u /mnt/hmc
```
hytop - Display hypervisor performance data

Use the `hytop` command to obtain a dynamic real-time view of a hypervisor environment on IBM Z.

It works with both the z/VM and the LPAR PR/SM™ hypervisor. Depending on the available data, it shows, for example, CPU and memory information about LPARs or z/VM guest virtual machines.

System names provided by hytop are either LPAR names as shown on the SE or HMC, or z/VM guest IDs that identify z/VM guest virtual machines.

The `hytop` command provides two main windows:
- A list of systems that the hypervisor is currently running (sys_list).
- One system in more detail (sys).

You can run `hytop` in interactive mode (default) or in batch mode with the `-b` option.

**Before you begin:**
- The debugfs file system must be mounted, see “debugfs” on page xi.
- The Linux kernel must have the required support to provide the performance data. Check that `<debugfs mount point>/s390_hypfs` is available after you mount debugfs.
- The hytop user must have read permission for the required debugfs files:
  - z/VM: `<debugfs mount point>/s390_hypfs/diag_2fc`
  - z/VM: `<debugfs mount point>/s390_hypfs/diag_0c` (Required only for management time data, identifiers \(m\) and \(M\). See “z/VM fields” on page 555)
  - LPAR: `<debugfs mount point>/s390_hypfs/diag_204`
- You can always monitor the guest operating system where hytop is running. To monitor any other operating system instances running on the same hypervisor as hytop, you will need additional permissions:
  - For z/VM: The guest virtual machine must be assigned privilege class B.
  - For LPAR: On the HMC or SE security menu of the LPAR activation profile, select the Global performance data control check box.
hyptop

hyptop syntax

```
hyptop [-w sys_list] [-w sys] [-s <system>]
        [-f <field>[:<unit>]] [-S <field>] [-t <type>] [-b] [-d <seconds>] [-n <iterations>]
```

Where:

- **-w <window name>** or **--window=<window name>**
  selects the window to display, either sys or sys_list. Use the options **--sys**, **--fields**, and **--sort** to modify the current window. The last window that is specified with the **--window** option is used as the start window. The default window is sys_list.

- **-s <system>** or **--sys=<system>**
  selects systems for the current window. If you specify this option, only the selected systems are shown in the window. For the sys window, you can specify only one system. <system> can be an LPAR name as shown on the SE or HMC, or it can be a z/VM guest ID that identifies a z/VM guest virtual machine. Enter hyptop without any options to display the names of all available systems.

- **-f <field>[:<unit>]** or **--fields=<field>[:<unit>]**
  selects fields and units in the current window. The <field> variable is a one letter unique identifier for a field (for example "c" for CPU time). The <unit> variable specifies the unit that is used for the field (for example “us” for microseconds). See "Available fields and units" on page 555 for definitions. If the **--fields** option is specified, only the selected fields are shown.

  **Note:** If your field specification includes the number sign (#), enclose the specification in double quotation marks. Otherwise, the command shell might interpret the number sign and all characters that follow as a comment.

- **-S <field>** or **--sort=<field>**
  selects the field that is used to sort the data in the current window. To reverse the sort order, specify the option twice. See "Available fields and units" on page 555 for definitions.

- **-t <type>** or **--cpu_types=<type>**
  selects CPU types that are used for dispatch time calculations. See "CPU types" on page 557 for definitions.
-b or --batch_mode
uses batch mode. Batch mode can be useful for sending output from hyptop to another program, a file, or a line mode terminal. In this mode, no user input is accepted.

-d <seconds> or --delay=<seconds>
specifies the delay between screen updates.

-n <iterations> or --iterations=<iterations>
specifies the maximum number of screen updates before the program ends.

-h or --help
prints usage information, then exits. To view the man page, enter man hyptop.

-v or --version
displays the version of hyptop, then exits.

Navigating between windows
Use letter or arrow keys to navigate between the windows.

When you start the hyptop command, the sys_list window opens in normal mode. Data is updated at regular intervals, and sorted by dispatch time. You can navigate between the windows as shown in Figure 81.

![Figure 81. hyptop window navigation overview](image)

To navigate between the windows, use the and arrow keys. The windows have two modes, normal mode and select mode.

You can get online help for every window by pressing the key. Press Q in the sys_list window to exit hyptop.

Instead of using the arrow keys, you can use letter keys (equivalent to the vi editor navigation) in all windows as listed in Table 60.

<table>
<thead>
<tr>
<th>Arrow key</th>
<th>Letter key equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>←</td>
<td>H</td>
</tr>
<tr>
<td>↓</td>
<td>J</td>
</tr>
<tr>
<td>↑</td>
<td>K</td>
</tr>
<tr>
<td>→</td>
<td>L</td>
</tr>
</tbody>
</table>

Selecting data
You can scroll windows and select data rows.
To enter select mode press the key. The display is frozen so that you can select rows. Select rows by pressing the and keys and mark the rows with the Spacebar. Marked rows are displayed in bold font. Leave the select mode by pressing the key.

To see the details of one system, enter select mode in the sys_list window, then navigate to the row for the system you want to look at, and press the key. The sys window for the system opens. The key always returns you to the previous window.

To scroll any window, press the and keys or the Page Up and Page Down keys. Jump to the end of a window by pressing the keys and to the beginning by pressing the key.

**Sorting data**

You can sort data according to column.

The sys window or sys_list window table is sorted according to the values in the selected column. Select a column by pressing the hot key of the column. This key is underlined in the heading. If you press the hot key again, the sort order is reversed. Alternatively, you can select columns with the and keys.

**Filtering data**

You can filter the displayed data by CPU types and by data fields.

From the sys or sys_list window you can access the fields selection window and the CPU-type selection window as shown in Figure 82.

![Diagram](image)

*Figure 82. Accessing the fields and CPU-type selection windows*

Use the key to toggle between the CPU-type selection window and the main window. Use the key to toggle between the fields selection window and the main window. You can also use the key to return to the main window from the CPU types and fields windows.

In the fields and CPU-type selection windows, press the field or CPU type identifier key (see “LPAR fields” on page 555, “z/VM fields” on page 555, and “CPU types” on page 557) to select or de-select. Selected rows are bold and de-selected rows are grey. When you return to the main window, the data is filtered according to your field and CPU type selections.
Available fields and units

Different fields are supported depending whether your hypervisor is LPAR PR/SM or z/VM.

The fields might also be different depending on machine type, z/VM version, and kernel version. Each field has a unique one letter identifier that can be used in interactive mode to enable the field in the field selection window. Also, use it to select the sort field in the sys or sys_list window. You can also select fields and sort data using the --fields and --sort command line options.

LPAR fields
Some fields for Linux in LPAR mode are available in both the sys_list and sys windows others are available only in the sys_list window or only in the sys window.

The following fields are available under LPAR in both the sys_list and sys windows:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Column label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>core</td>
<td>Core dispatch time per second</td>
</tr>
<tr>
<td>e</td>
<td>teh</td>
<td>Thread time per second</td>
</tr>
<tr>
<td>m</td>
<td>mgm</td>
<td>Management time per second</td>
</tr>
<tr>
<td>C</td>
<td>Core+</td>
<td>Total core dispatch time</td>
</tr>
<tr>
<td>E</td>
<td>thE+</td>
<td>Total thread time</td>
</tr>
<tr>
<td>M</td>
<td>Mgm+</td>
<td>Total management time</td>
</tr>
<tr>
<td>o</td>
<td>online</td>
<td>Online time</td>
</tr>
</tbody>
</table>

If multithreading is not available or not enabled, the values for core and for thread are identical.

In the sys_list window only:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Column label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>system</td>
<td>Name of the LPAR (always shown)</td>
</tr>
<tr>
<td>#</td>
<td>#cpu</td>
<td>Number of CPUs</td>
</tr>
<tr>
<td>T</td>
<td>#The</td>
<td>Number of threads (sum of initial and reserved)</td>
</tr>
</tbody>
</table>

In the sys window only:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Column label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>coreid</td>
<td>Core identifier (always shown)</td>
</tr>
<tr>
<td>p</td>
<td>type</td>
<td>CPU type. See “CPU types” on page 557</td>
</tr>
<tr>
<td>v</td>
<td>visual</td>
<td>Visualization of core dispatch time per second</td>
</tr>
</tbody>
</table>

z/VM fields
Some fields for Linux on z/VM are available in both the sys_list and sys windows. Others are available only in the sys_list window or only in the sys window.

In the sys_list and sys windows:
<table>
<thead>
<tr>
<th>Identifier</th>
<th>Column label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>cpu</td>
<td>CPU time per second</td>
</tr>
<tr>
<td>m</td>
<td>mgm</td>
<td>Management time per second</td>
</tr>
<tr>
<td>C</td>
<td>Cpu+</td>
<td>Total CPU time</td>
</tr>
<tr>
<td>M</td>
<td>Mgm+</td>
<td>Total management time</td>
</tr>
<tr>
<td>o</td>
<td>online</td>
<td>Online time</td>
</tr>
</tbody>
</table>

In the `sys_list` window only:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Column label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>system</td>
<td>Name of the z/VM guest virtual machine (always shown)</td>
</tr>
<tr>
<td>#</td>
<td>#cpu</td>
<td>Number of CPUs</td>
</tr>
<tr>
<td>O</td>
<td>#cpuop</td>
<td>Number of operating CPUs</td>
</tr>
<tr>
<td>u</td>
<td>memuse</td>
<td>Used memory</td>
</tr>
<tr>
<td>a</td>
<td>memmax</td>
<td>Maximum memory</td>
</tr>
<tr>
<td>r</td>
<td>wcmax</td>
<td>Current weight</td>
</tr>
<tr>
<td>x</td>
<td>wmax</td>
<td>Maximum weight</td>
</tr>
</tbody>
</table>

In the `sys` window only:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Column label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>cpuid</td>
<td>CPU identifier (always shown)</td>
</tr>
<tr>
<td>v</td>
<td>visual</td>
<td>Visualization of CPU time per second</td>
</tr>
</tbody>
</table>

**Units**

Depending on the field type, the values can be displayed in different units.

In the `sys_list` and `sys` windows, the units are displayed under the column headings in parenthesis. Each unit can be specified through the `--fields` command line option. Units can also be selected interactively. To change a unit, enter select mode in the fields window. Then, select the field where you want to change the unit, and press the "+" or "-" keys to go through the available units. The following units are supported:

**Units of time:**

<table>
<thead>
<tr>
<th>Unit</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>us</td>
<td>Microseconds ($10^{-6}$ seconds)</td>
</tr>
<tr>
<td>ms</td>
<td>Milliseconds ($10^{-3}$ seconds)</td>
</tr>
<tr>
<td>%</td>
<td>Hundreds of a second ($10^{-2}$ seconds) or percent</td>
</tr>
<tr>
<td>s</td>
<td>Seconds</td>
</tr>
<tr>
<td>m</td>
<td>Minutes</td>
</tr>
<tr>
<td>hm</td>
<td>Hours and minutes</td>
</tr>
<tr>
<td>dhm</td>
<td>Days, hours, and minutes</td>
</tr>
</tbody>
</table>
Units of memory:

<table>
<thead>
<tr>
<th>Unit</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>KiB</td>
<td>Kibibytes (1,024 bytes)</td>
</tr>
<tr>
<td>MiB</td>
<td>Mebibytes (1,048,576 bytes)</td>
</tr>
<tr>
<td>GiB</td>
<td>Gibibytes (1,073,741,824 bytes)</td>
</tr>
</tbody>
</table>

Other units:

<table>
<thead>
<tr>
<th>Unit</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>str</td>
<td>String</td>
</tr>
<tr>
<td>#</td>
<td>Count or number</td>
</tr>
<tr>
<td>vis</td>
<td>Visualization</td>
</tr>
</tbody>
</table>

**CPU types**

Enable or disable CPU types in interactive mode in the cpu_types window.

The CPU types can also be specified with the `--cpu_types` command line option.

The calculation of the CPU data uses CPUs of the specified types only. For example, if you want to see how much CPU time is consumed by your Linux systems, enable CPU type IFL.

On z/VM the processor type is always UN and you cannot select the type.

In an LPAR the following CPU types can be selected either interactively or with the `--cpu_types` command line option:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Column label</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>IFL</td>
<td>Integrated Facility for Linux. On older machines IFLs might be shown as CPs.</td>
</tr>
<tr>
<td>p</td>
<td>CP</td>
<td>CP processor type.</td>
</tr>
<tr>
<td>u</td>
<td>UN</td>
<td>Unspecified processor type (other than CP or IFL).</td>
</tr>
</tbody>
</table>

**Examples**

These examples show typical uses of `hyptop`.

- To start `hyptop` with the `sys_list` window in interactive mode, enter:
  ```
  # hyptop
  ```

- If your Linux instance is running in an LPAR that has permission to see the other LPARs, the output looks like the following example:
- If your Linux instance runs in a z/VM guest virtual machine that has permission to see the other z/VM guest virtual machines, the output looks like the following example:

```
12:30:48 | cpu-t: IFL(18) CP(3) UN(3) ?=help
system core mgm core+ mgm+ online
1 (st) (%) (%) (Tm) (Thm) (dhm)
S05LP30 10 461.14 10.18 1547:41 8:15 11:05:59
S05LP33 4 133.73 7.57 220:53 6:12 11:05:54
S05LP50 4 99.26 0.01 146:24 0:12 10:04:24
S05LP02 1 99.09 0.00 269:57 0:00 11:05:58
TRX2CEFA 1 2.74 0.03 3:24 0:04 11:06:01
S05LP13 6 1.36 0.34 4:23 0:54 11:05:56
TRX1 19 1.77 0.14 13:57 0:22 11:06:01
TRX2 20 1.76 0.11 26:05 0:25 11:06:00
S05LP55 2 0.00 0.00 0:00 0:00 11:05:52
S05LP56 3 0.00 0.00 0:00 0:00 11:05:52
413 823.39 23.86 3159:57 38:08 11:06:01
```

At the top, the sys and sys_list windows show a list of the CPU types that are used for the current CPU and core dispatch time calculation.

- To start `hyptop` with the sys window showing performance data for LPAR MYLPAR, enter:

  ```
  # hyptop -w sys -s mylpar
  ```

The result looks like the following example:

```
12:32:21 | CPU-T: UN(16) ?=help
system #cpu cpu Cpu+ online memuse memmax wcur
(core) (mem) (mem) (core) (vis)
1 100.31 53:05:20 1.56 2.00 100
2 0.00 0:00 0.01 0.03 100
3 0.00 0:37 0.32 0.75 100
4 0.00 3700:57 4.00 4.00 100
5 0.00 0:00 0.00 0.03 100
6 0.00 0:00 0.00 0.02 100
7 0.00 0:00 0.00 0.01 100
8 0.00 0:00 0.00 0.01 100
9 0.00 0:00 0.00 0.01 100
10 0.00 0:00 0.00 0.01 100
11 0.00 0:00 0.00 0.01 100
12 0.00 0:00 0.00 0.01 100
13 0.00 0:00 0.00 0.01 100
14 0.00 0:00 0.00 0.01 100
15 0.00 0:00 0.00 0.01 100
16 0.00 0:00 0.00 0.01 100
```

- To start `hyptop` with the sys_list window in batch mode, enter:

  ```
  # hyptop
  ```
To start **hytop** with the sys_list window in interactive mode, with the fields dispatch time (in milliseconds), and online time (unit default), and sort the output according to online time, enter:

```
# hytop -f c:ms,o -S o
```

To start **hytop** with the sys_list window in batch mode with update delay 5 seconds and 10 iterations, enter:

```
# hytop -b -d 5 -n 10
```

To start **hytop** with the sys_list window and use only CPU types IFL and CP for dispatch time calculation, enter:

```
# hytop -t ifl,cp
```

To start **hytop** on Linux in LPAR mode with the sys_list window and display all LPAR fields, including the thread information, enter:

```
# hytop -f "#,T,c,e,m,C,E,M,o"
```

The result looks like the following example:

```
13:47:42 cpu-t: IFL(0) CP(38) UN(0) ?=help
system  #core  #The  core  the  mgm  Core+  thE+  Mgm+  onTime
         (#)   (%)   (%)   (%)  (%)   (%)   (%)  (%)   (hm)  (hm)  (hm)  (dhm)
S35LP41  12   24  101.28  170.28  0.28  1056:10  1756:11  8:45  158:04:04
S35LP42  16   32  35.07   40.07  0.44  5194:52  6193:52 12:45  158:04:04
S35LP64  3   3   1.20   1.20   0.00    0:31    0:31    0:00  12:03:54
...
```

In the example, the Linux instances in LPARs S35LP41 and S35LP43 run with 2 threads per core. The thread time, as the sum of the two threads, exceeds the core dispatch time.

The Linux instance in LPAR S35LP64 does not use simultaneous multithreading.

To start **hytop** on Linux on z/VM with the sys_list window and display a selection of z/VM fields, including the management time, enter:

```
# hytop -f "#,c,m,C,M,o"
```

The result looks like the following example:

```
17:52:56 cpu-t: IFL(0) UN(2) ?=help
system    #cpu  cpu  mgm  Cpu+  Mgm+  online
         (#)  (%)  (%)  (%)  (%)  (dhm)
G3545010  3  0.55  0.05  0:05  0:02  0:03:14
G3545021  3  0.04  -0:00  -0:02  0:02:43
G3545025  2  0.01  -0:00  -0:04  0:04:08
...
G3545099  1  0.00  -0:00  -0:09 0:09:06
... 52  0.61  0.05  0:27  0:02  0:09:06
```

In the example, hytop runs on a Linux instance in z/VM guest virtual machine G3545010. In the sys_list window, this is the only guest virtual machine for which management data is displayed.
Scenario
Perform the steps described in this scenario to start hyptop with the sys window with system MYLPAR with the fields dispatch time (unit milliseconds) and total dispatch time (unit default), sort the output according to the total dispatch time, and then reverse the sort order.

Procedure
1. Start hyptop.
   
   # hyptop

2. Go to select mode by pressing the key. The display will freeze.

3. Navigate to the row for the system you want to look (in the example MYLPAR) at using the and keys.

   12:15:00 CPU-T: IFL(18) CP(3) UN(3) 
   system #core core mgm Core+ Mgm+ online
   (str) (s) (s) (s) (s) (s) 
   MYLPAR 199.09 0.04 54:41:00 11:05:19
   S05LP33 4 133.73 7.57 220:53 6:12 11:05:54
   S05LP50 4 99.26 0.01 146:24 0:12 10:04:24
   S05LP02 1 99.09 0.00 269:57 0:00 11:05:58
   ...
   S05LP56 3 0.00 0.00 0:00 0:00 11:05:52
   413 823.39 23.86 3159:57 38:08 11:06:01

4. Open the sys window for MYLPAR by pressing the key.

   12:15:51 MYLPAR CPU-T: IFL(18) CP(3) UN(2)
   coreid type core mgm visual
   (#) (str) (%) (%) (vis)
   0 IFL 99.84 0.02
   1 IFL 99.85 0.02
   2 IFL 0.00 0.00
   3 IFL 0.00 0.00
   :V:N 199.69 0.04

5. Press the key to go to the fields selection window:

   Select Fields and Units
   # S ID UNIT AGG DESCRIPTION 
   p * type str none CPU type
   c * core % sum Core dispatch time per second
   e the % sum Thread time per second
   m * mgm % sum Management time per second
   C core+ hm sum Total core dispatch time
   E the+ % sum Total thread time
   M mgm+ hm sum Total management time
   o online dhm max Online time
   v * visual vis none Visualization of CPU time per second

Ensure that dispatch time per second and total dispatch time are selected and for dispatch time microseconds are used as unit:

a. Press the key, the key, and the key to disable CPU type, Management time per second, and Visualization.

b. Press the key to enable Total core dispatch time.

c. Then select the Core dispatch time per second row by pressing the and keys.
d. Press the minus key (-) to switch from the percentage (%) unit to the microseconds (ms) unit.

Press the minus key (-) to switch from the percentage (%) unit to the microseconds (ms) unit.

<table>
<thead>
<tr>
<th>Select Fields and Units</th>
<th>?=help</th>
</tr>
</thead>
<tbody>
<tr>
<td>K S ID</td>
<td>UNIT AGG DESCRIPTION</td>
</tr>
<tr>
<td>p type str</td>
<td>none CPU type</td>
</tr>
<tr>
<td>e * core ms sum</td>
<td>Core dispatch time per second</td>
</tr>
<tr>
<td>e the ¾ sum</td>
<td>Thread time per second</td>
</tr>
<tr>
<td>m mgm ¾ sum</td>
<td>Management time per second</td>
</tr>
<tr>
<td>C * core+ hm sum</td>
<td>Total core dispatch time</td>
</tr>
<tr>
<td>E theE+ ¾ sum</td>
<td>Total thread time</td>
</tr>
<tr>
<td>M mgm+ hm sum</td>
<td>Total management time</td>
</tr>
<tr>
<td>o online dhm max</td>
<td>Online time</td>
</tr>
<tr>
<td>v visual vis</td>
<td>none Visualization of CPU time per second</td>
</tr>
</tbody>
</table>

Press the right arrow key twice to return to the sys window.

6. To sort by Total core dispatch time and list the values from low to high, press the Shift + C keys twice:

<table>
<thead>
<tr>
<th>13:44:41 MYPAR cpu-t:</th>
<th>CPU(18) CP(3) UN(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>coreid core Core+ (s)</td>
<td>(ms) (hm)</td>
</tr>
<tr>
<td>2</td>
<td>0.00 0:00</td>
</tr>
<tr>
<td>3</td>
<td>0.00 0:00</td>
</tr>
<tr>
<td>1</td>
<td>37.48 492:55</td>
</tr>
<tr>
<td>0</td>
<td>23.84 548:52</td>
</tr>
<tr>
<td>:^N 61.33 1041:47</td>
<td></td>
</tr>
</tbody>
</table>

Results

You can do all of these steps in one by entering the command:

```bash
# hyptop -w sys -s mylpar -f c:ms,C -S C -S C
```
lschp

lschp - List channel paths

Use the lschp command to display information about channel paths.

lschp syntax

```
lschp
  --help
  --version
```

Where:

- **-v or --version**
  displays the version number of lschp and exits.

- **-h or --help**
  displays a short help text, then exits.

Output column description:

- **CHPID**
  Channel-path identifier.

- **Vary**
  Logical channel-path state:
  - 0 = channel-path is not used for I/O.
  - 1 = channel-path is used for I/O.

- **Cfg.**
  Channel-path configure state:
  - 0 = stand-by
  - 1 = configured
  - 2 = reserved
  - 3 = not recognized

- **Type**
  Channel-path type identifier.

- **Cmg**
  Channel measurement group identifier.

- **Shared**
  Indicates whether a channel-path is shared between LPARs:
  - 0 = channel-path is not shared
  - 1 = channel-path is shared

- **PCHID**
  Physical channel path identifier, or, if enclosed in brackets, internal channel identifier. The mapping might not be available to Linux when running as a z/VM guest. If so, use the CP command:
  ```
  QUERY CHPID <num> PCHID
  ```

A column value of '-' indicates that a facility associated with the corresponding channel-path attribute is not available.
Examples

- To query the configuration status of channel path ID 0.40 issue:

```
# lschp

CHPID Vary Cfg. Type Cmg Shared PCHID
======================================
...                                      
...                                      
0.40 1 1 1b 2 1 0580                     
...                                      
```

The value under Cfg. shows that the channel path is configured (1).
Isscpumf - Display information about the CPU-measurement facilities

Use the `lsscpumf` command to display information about information about the CPU-measurement facilities.

**lsscpumf syntax**

```
lsscpumf
  -i
  -c
  -C
  -s
  -h
  -v
```

where:

- **-i** or **--info**
  displays detailed information about available and supported CPU measurement facilities.

- **-c** or **--list-counters**
  lists counters that are provided by the CPU-measurement facility, omitting counters for which the LPAR is not authorized. For counter measurements with the `perf` program, the raw event identifier and symbolic counter name are displayed.

- **-C** or **--list-all-counters**
  lists all counters that are provided by the CPU-measurement counter facility, regardless of LPAR authorization. To list only those counters for which the LPAR is authorized, use the **-c** option. For counter measurements with the `perf` program, the raw event identifier and symbolic counter name are displayed.

- **-s** or **--list-sampling-events**
  lists `perf` raw events that activate the sampling facility.

- **-v** or **--version**
  displays the version number of `lsscpumf` and exits.

- **-h** or **--help**
  displays out a short help text, then exits. To view the man page, enter `man lscpumf`.

**Examples**

- To display the supported facilities, issue:

  ```
  # lscpumf
  CPU-measurement Counter Facility
  CPU-measurement Sampling Facility
  ```

- To display details about the facilities, issue:
To display perf event information for authorized sampling functions, issue:

```
# lscpumf -s
```

<table>
<thead>
<tr>
<th>Raw event Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF_CYCLES_BASIC</td>
<td>Sample CPU cycles using basic-sampling mode.</td>
</tr>
<tr>
<td></td>
<td>This event is not associated with a counter set.</td>
</tr>
</tbody>
</table>

To list all counters that are provided by your IBM Z hardware, issue:
<table>
<thead>
<tr>
<th>Raw event</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>r0</td>
<td>CPU_CYCLES</td>
<td>Cycle Count. Counter 0 / Basic Counter Set.</td>
</tr>
<tr>
<td>r1</td>
<td>INSTRUCTIONS</td>
<td>Instruction Count. Counter 1 / Basic Counter Set.</td>
</tr>
<tr>
<td>r2</td>
<td>L1I_DIR_WRITES</td>
<td>Level-1 I-Cache Directory Write Count. Counter 2 / Basic Counter Set.</td>
</tr>
<tr>
<td>r3</td>
<td>L1I_PENALTY_CYCLES</td>
<td>Level-1 I-Cache Penalty Cycle Count. Counter 3 / Basic Counter Set.</td>
</tr>
<tr>
<td>r4</td>
<td>L1D_DIR_WRITES</td>
<td>Level-1 D-Cache Directory Write Count. Counter 4 / Basic Counter Set.</td>
</tr>
<tr>
<td>r5</td>
<td>L1D_PENALTY_CYCLES</td>
<td>Level-1 D-Cache Penalty Cycle Count. Counter 5 / Basic Counter Set.</td>
</tr>
<tr>
<td>r20</td>
<td>PROBLEM_STATE_CPU_CYCLES</td>
<td>Problem-State Cycle Count. Counter 32 / Problem-State Counter Set.</td>
</tr>
<tr>
<td>r21</td>
<td>PROBLEM_STATE_INSTRUCTIONS</td>
<td>Problem-State Instruction Count. Counter 33 / Problem-State Counter Set.</td>
</tr>
</tbody>
</table>

...
Iscss - List subchannels

Use the `lscss` command to gather subchannel information from sysfs and display it in a summary format.

### lscss syntax

```
lscss [options] [device]
```

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>-s</code></td>
<td>strips the bus ID in the command output down to the four-digit device number.</td>
</tr>
<tr>
<td><code>-u</code></td>
<td>displays the output with uppercase letters. The default is lowercase.</td>
</tr>
<tr>
<td><code>--avail</code></td>
<td>includes the availability attribute of I/O devices.</td>
</tr>
<tr>
<td><code>--vpm</code></td>
<td>shows verified paths in a mask. Channel paths listed in this mask are available to Linux device drivers for I/O. Reasons for a channel path being unavailable include:</td>
</tr>
<tr>
<td></td>
<td>• The corresponding bit is not set in at least one of the PIM, PAM, or POM masks.</td>
</tr>
<tr>
<td></td>
<td>• The channel path is varied offline.</td>
</tr>
<tr>
<td></td>
<td>• Linux received no interrupt to I/O using this channel path.</td>
</tr>
<tr>
<td><code>--io</code></td>
<td>limits the output to I/O subchannels and corresponding devices. This option is the default.</td>
</tr>
</tbody>
</table>

**Changed default:** Earlier versions of `lscss` printed the command output in uppercase. Specify this option to obtain the former output style.
\textbf{lscss}

\textit{--chsc}
limits the output to CHSC subchannels.

\textit{--eadm}
limits the output to EADM subchannels.

\textit{-a or --all}
does not limit the output.

\textit{-t or --devtype}
limits the output to information about the specified device types and, if provided, the specified model.

\textit{<devicetype>}
specifies a device type.

\textit{<model>}
is a specific model of the specified device type.

\textit{-d or --devrange}
interprets bus IDs as specifications of devices. By default, bus IDs are interpreted as specifications of subchannels.

\textit{<bus_id>}
specifies an individual subchannel; if used with -d specifies an individual device. If you omit the leading 0.<subchannel set ID>, 0.0. is assumed.

If you specify subchannels or devices, the command output is limited to these subchannels or devices.

\textit{<from_bus_id>-<to_bus_id>}
specifies a range of subchannels; if used with -d specifies a range of devices. If you omit the leading 0.<subchannel set ID>, 0.0. is assumed.

If you specify subchannels or devices, the command output is limited to these subchannels or devices.

\textit{-v or --version}
displays the version number of \texttt{lscss} and exits.

\textit{-h or --help}
displays a short help text, then exits. To view the man page enter \texttt{man lscss}.

\textbf{Examples}

\begin{itemize}
  \item This command lists all subchannels, including subchannels that do not correspond to I/O devices:
\end{itemize}
This command lists subchannels with an attached 3480 model 04 or 3590 tape device and strips the bus ID and the subchannel ID in the command output down to the four-digit IDs:

```
# lscss -a -s -t 3480/04,3590
Device Subchan. DevType CU Type Use PIM PAM POM CHPID
0480 000e 3480/04 3480/01 80 80 ff 10000000 00000000
0a38 000f 3590/11 3590/50 80 80 fc 34400000 00000000
```

This command limits the output to subchannels with attached DASD model 3390 type 0a:

```
# lscss -t 3390/0a
Device Subchan. DevType CU Type Use PIM PAM POM CHPID
0.0.2f08 0.0.0a78 3390/0a 3990/e9 yes c0 c0 ff 34400000 00000000
0.0.2fe7 0.0.0b57 3390/0a 3990/e9 yes c0 c0 ff 34400000 00000000
```

This command limits the output to the subchannel range 0.0.0b00-0.0.0bff:

```
# lscss 0.0.0b00-0.0.0bff
Device Subchan. DevType CU Type Use PIM PAM POM CHPID
0.0.2fe7 0.0.0b57 3390/0a 3990/e9 yes c0 c0 ff 34400000 00000000
```

This command limits the output to subchannels 0.0.0a78 and 0.0.0b57 and shows the availability:

```
# lscss --avail 0a78,0b57
Device Subchan. DevType CU Type Use PIM PAM POM CHPID Avail.
0.0.2f08 0.0.0a78 3390/0a 3990/e9 yes c0 c0 ff 34400000 00000000 good
0.0.2fe7 0.0.0b57 3390/0a 3990/e9 yes c0 c0 ff 34400000 00000000 good
```
This command limits the output to subchannel 0.0.0a78 and displays uppercase output:

```
# lscss -u 0a78
Device Subchan. DevType CU Type Use PIM PAM POM CHPIDs
0.0.2f08 0.0.0a78 3390/0a 3990/e9 YES CO CO FF 34400000 00000000
```

This command limits the output to subchannels that correspond to I/O device 0.0.7e10 and the device range 0.0.2f00-0.2fff:

```
# lscss -d 2f00-2fff,0.0.7e10
Device Subchan. DevType CU Type Use PIM PAM POM CHPIDs
0.0.2f08 0.0.0a78 3390/0a 3990/e9 yes c0 c0 ff 34400000 00000000
0.0.2fe5 0.0.0b55 3390/0a 3990/e9 c0 c0 bf 34400000 00000000
0.0.2fe6 0.0.0b56 3390/0a 3990/e9 c0 c0 bf 34400000 00000000
0.0.2fe7 0.0.0b57 3390/0a 3990/e9 yes c0 c0 ff 34400000 00000000
0.0.7e10 0.0.1828 3390/0c 3990/e9 yes f0 f0 ef 34403541 00000000
```

This example shows a CHPID with PIM, PAM, and POM masks that are OK, but the entry in the VPM column indicates that one of the paths, 0x41, is not usable for I/O.

```
# lscss --vpm
Device Subchan. DevType CU Type Use PIM PAM POM VPM CHPIDs
0.0.f500 0.0.05cf 1732/01 1731/01 yes 80 80 ff 80 76000000 00000000
0.0.f501 0.0.05d0 1732/01 1731/01 yes 80 80 ff 80 76000000 00000000
0.0.f502 0.0.05d1 1732/01 1731/01 yes 80 80 ff 80 76000000 00000000
0.0.6194 0.0.3700 3390/0c 3990/e9 yes fc fc ff f8 32333435 40410000
0.0.6195 0.0.3701 3390/0c 3990/e9 yes fc fc ff f8 32333435 40410000
0.0.6196 0.0.3702 3390/0c 3990/e9 yes fc fc ff f8 32333435 40410000
0.0.6197 0.0.3703 3390/0c 3990/e9 fc fc ff 00 32333435 40410000
0.2.5600 0.2.0040 1732/03 1731/03 80 80 ff 00 5d000000 00000000
```
Isdasd - List DASD devices

Use the `lsdasd` command to gather information about DASD devices from sysfs and display it in a summary format.

Isdasd syntax

```
```

Where:

- `-a` or `--offline`
  includes devices that are currently offline.

- `-b` or `--base`
  omits PAV alias devices. Lists only base devices.

- `-s` or `--short`
  strips the bus ID in the command output down to the four-digit device number.

- `-v` or `--verbose`
  Obsolete. This option has no effect on the output.

- `-l` or `--long`
  extends the output to include attributes, the UID and path information.

- `-c` or `--compat`
  creates output of this command as with versions earlier than 1.7.0.

- `-u` or `--uid`
  includes and sorts output by UID.

- `-H` or `--host_access_list`
  shows information about all operating system instances that use this device.

- `--version`
  displays the version of the command.

- `<device_bus_id>`
  limits the output to information about the specified devices only.

- `-h` or `--help`
  displays a short help text, then exits.
Examples

- The following command lists all DASD (including offline DASDs):

```bash
# lsdasd -a

<table>
<thead>
<tr>
<th>Bus-ID</th>
<th>Status</th>
<th>Name</th>
<th>Device</th>
<th>Type</th>
<th>BlkSz</th>
<th>Size</th>
<th>Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.0190</td>
<td>offline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0.0191</td>
<td>offline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0.019d</td>
<td>offline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0.019e</td>
<td>offline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0.0592</td>
<td>offline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0.4711</td>
<td>offline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0.4712</td>
<td>offline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0.4f2c</td>
<td>offline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0.4d80</td>
<td>active</td>
<td>dasda</td>
<td>94:0</td>
<td>ECKD</td>
<td>4096</td>
<td>4695MB</td>
<td>1202040</td>
</tr>
<tr>
<td>0.0.4f19</td>
<td>active</td>
<td>dasdb</td>
<td>94:4</td>
<td>ECKD</td>
<td>4096</td>
<td>23034MB</td>
<td>5896800</td>
</tr>
<tr>
<td>0.0.4d81</td>
<td>active</td>
<td>dasdc</td>
<td>94:8</td>
<td>ECKD</td>
<td>4096</td>
<td>4695MB</td>
<td>1202040</td>
</tr>
<tr>
<td>0.0.4d82</td>
<td>active</td>
<td>dasdd</td>
<td>94:12</td>
<td>ECKD</td>
<td>4096</td>
<td>4695MB</td>
<td>1202040</td>
</tr>
<tr>
<td>0.0.4d83</td>
<td>active</td>
<td>dasde</td>
<td>94:16</td>
<td>ECKD</td>
<td>4096</td>
<td>4695MB</td>
<td>1202040</td>
</tr>
</tbody>
</table>
```

- The following command shows information only for the DASD with device number 0x4d80 and strips the bus ID in the command output down to the device number:

```bash
# lsdasd -s 0.0.4d80

<table>
<thead>
<tr>
<th>Bus-ID</th>
<th>Status</th>
<th>Name</th>
<th>Device</th>
<th>Type</th>
<th>BlkSz</th>
<th>Size</th>
<th>Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>4d80</td>
<td>active</td>
<td>dasda</td>
<td>94:0</td>
<td>ECKD</td>
<td>4096</td>
<td>4695MB</td>
<td>1202040</td>
</tr>
</tbody>
</table>
```

- The following command shows only online DASDs in the format of `lsdasd` versions earlier than 1.7.0:

```bash
# lsdasd -c

0.0.4d80(ECKD) at { 94: 0 } is dasda: active at blocksize 4096, 1202040 blocks, 4695 MB
0.0.4f19(ECKD) at { 94: 4 } is dasdb: active at blocksize 4096, 5896800 blocks, 23034 MB
0.0.4d81(ECKD) at { 94: 8 } is dasdc: active at blocksize 4096, 1202040 blocks, 4695 MB
0.0.4d82(ECKD) at { 94: 12 } is dasdd: active at blocksize 4096, 1202040 blocks, 4695 MB
0.0.4d83(ECKD) at { 94: 16 } is dasde: active at blocksize 4096, 1202040 blocks, 4695 MB
```

- The following command shows the device geometry, UID, path information, and some of the settings for the DASD with device bus-ID 0.0.4d82:

```bash
lsdasd -l 0.0.4d82

0.0.4d82/dasdd/94:12

status: active

<table>
<thead>
<tr>
<th>option</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td>ECKD</td>
</tr>
<tr>
<td>blksz</td>
<td>4096</td>
</tr>
<tr>
<td>size</td>
<td>4695MB</td>
</tr>
<tr>
<td>blocks</td>
<td>1202040</td>
</tr>
<tr>
<td>use_diag</td>
<td>0</td>
</tr>
<tr>
<td>readonly</td>
<td>0</td>
</tr>
<tr>
<td>eer_enabled</td>
<td>0</td>
</tr>
<tr>
<td>erplog</td>
<td>0</td>
</tr>
<tr>
<td>hpf</td>
<td>1</td>
</tr>
<tr>
<td>uid</td>
<td>IBM.750000000010671.4d82.16</td>
</tr>
<tr>
<td>paths_installed</td>
<td>30 31 32 33 3c 3d</td>
</tr>
<tr>
<td>paths_in_use</td>
<td>31 32 33</td>
</tr>
<tr>
<td>paths_non_preferred</td>
<td>paths_invalid_cabling: 3c</td>
</tr>
<tr>
<td></td>
<td>paths_cuir_quiesced: 30</td>
</tr>
<tr>
<td></td>
<td>paths_invalid_hpf_characteristics: 3d</td>
</tr>
<tr>
<td></td>
<td>paths_error_threshold_exceeded:</td>
</tr>
</tbody>
</table>
```

In the example, three of the installed paths are unused for different reasons:
The path with CHPID 3c is not used because of a cabling error to the storage system. This channel path does not connect to the same physical disk space as the other channel path for this device.

The path with CHPID 30 is not used because of a control-unit initiated reconfiguration (CUIR).

The path with CHPID 3d is not used because its High Performance FICON characteristics do not match with the paths currently in use.

The following command shows whether other operating system instances access device 0.0.bf45:

```
# lsdasd -H bf45
```

<table>
<thead>
<tr>
<th>Path-Group-ID</th>
<th>LPAR</th>
<th>CPU</th>
<th>FL</th>
<th>Status</th>
<th>Sysplex</th>
<th>Max_Cyls</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>88000d29e72964ce8570b8 0d</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>TRX1LNX1</td>
<td>268434453</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>88000e29e72964ce8570c3 0e</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88000ff29e72964ce8570d1 0f</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88011d29e72964ce8570d4 1d</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88011e29e72964ce8570d9 1e</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88011f29e72964ce8570e3 1f</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88022d29e72964ce8570e6 2d</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88022e29e72964ce8570ea 2e</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88022f29e72964ce8570f1 2f</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88033d29e72964ce8570f7 3d</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88033e29e72964ce8570fe 3e</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88033f29e72964ce85710e 3f</td>
<td>29e7</td>
<td>50</td>
<td>ON</td>
<td>268434453</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80004229e72964ce7dce74 42</td>
<td>29e7</td>
<td>00</td>
<td>OFF</td>
<td>65520</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80004a29e72964ce7d6b6d 4a</td>
<td>29e7</td>
<td>00</td>
<td>OFF</td>
<td>65520</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80003c29e72964ce8481a6 3c</td>
<td>29e7</td>
<td>00</td>
<td>OFF</td>
<td>65520</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80004629e72964ce7f1c13 46</td>
<td>29e7</td>
<td>70</td>
<td>ON-RSV</td>
<td>65520</td>
<td>1424174863</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Status values are:

**ON** The device is online.

**OFF** The device is offline.

**ON-RSV** The device is online and reserved.

**OFF-RSV** The device is offline and reserved by an operating system instance in another LPAR.

The meaning of the columns is as follows:

**Path-group-ID**
A 22-digit hexadecimal number assigned by the operating system when setting the DASD online. This ID uniquely identifies the operating system to the storage server.

**LPAR** A 2 digit LPAR ID.

**CPU** A 4 digit CPU ID, as it is defined in the HMC or can be read from `/proc/cpuinfo`.

**FL** A 2 digit hexadecimal flag. 0x20 means reserved, 0x50 means online.

**Sysplex**
The 8-character EBCDIC name of the SYSPLEX.

**MAX_CYLS**
The maximum number of cylinders per volume that are supported by the host.
Isdasd

TIME  Time the device has been reserved in seconds since July 1, 1970.
Ishmc - List media contents in the HMC media drive

Use the `ishmc` command to display the contents of the media in the HMC media drive.

**Before you begin:** To be able to use this command, you need the `hmcdrv` module (see Chapter 27, “HMC media device driver,” on page 355).

**Ishmc syntax**

```
llshmc <filepath> -s
```

Where:

- `<filepath>` specifies a directory or path to a file to be listed. Path specifications are relative to the root of the file system on the media. You can use the asterisk (*) and question mark (?) as wildcards. If this specification is omitted, the contents of the root directory are listed.

- `-s` or `--short` limits the output to regular files in a short listing format. Omits directories, symbolic links, and device nodes and other special files.

- `-v` or `--version` displays version information for the command.

- `-h` or `--help` displays a short help text, then exits. To view the man page, enter `man lshmc`.

**Examples**

- To list the files in the root directory of the media in the HMC’s media drive, issue:
  ```
  # lshmc
  ```

- If the `hmcdrv` kernel module is not loaded, load it before you issue the `lshmc` command:
  ```
  # modprobe hmcdrv
  # lshmc
  ```

- To list all HTML files in subdirectory `www`, issue:
  ```
  # lshmc /www/*.
  ```
lsluns - Discover LUNs, or show encryption state of attached LUNs

Use the lsluns command to list logical unit numbers (LUNs) discovered in the Fibre Channel storage area networks (SAN), or to show the encryption state of zfcp-attached LUNs.

lsluns is designed for environments where all SCSI devices are attached through the zfcp device driver. Expect error messages in mixed environments such as with iSCSI.

lsluns lists all LUNs discovered in the Fibre Channel SAN. This discovery causes extra SAN traffic for each target port WWPN.

lsluns -a shows the encryption state of the attached LUNs. Running lsluns -a causes extra SAN traffic for each attached LUN.

Filter the listing by specifying one or more FCP device bus-IDs, target port WWPNs, or both.

For all other uses, such as listing attached LUNs or properties other than encryption, use other tools such as:

- lszfcp -D See “lszfcp - List zfcp devices” on page 595
- lsscsi -tvxx See the man page for more details.

Details on lsluns without the -a option

Prerequisite

Discovering LUNs only makes sense for NPIV-enabled FCP devices without zfcp automatic LUN scan. zfcp automatic LUN scan is available as of kernel version 2.6.37, if not disabled with zfcp.allow_lun_scan=0. See “Setting up the zfcp device driver” on page 165.

With available and enabled zfcp automatic LUN scan, the kernel performs LUN discovery.

Temporary LUN Attachment

If not attached already, lsluns temporarily attaches LUN 0 (or if this fails, the WLUN 0xc101000000000000) during runtime. Do not terminate lsluns with a signal. Signals interfere with the removal of temporarily attached LUNs.

Storage Products

Some storage products return a peripheral device type of 31==0x1f with peripheral qualifier 0 in a SCSI standard INQUIRY command for an unmapped FCP LUN 0. Examples are: IBM Storwize products, including IBM V7000, IBM V840, IBM V9000, and IBM SAN Volume Controller. For lsluns to work with such storage products, you must have a host mapping on the storage side, which maps some volume to exported FCP LUN 0x0000000000000000 (Storwize host map property "SCSI ID" 0) for each used FCP-device initiator WWPN. The volume can be a minimum-sized thin-provisioned shared stand-in volume.
Islns syntax

Where:

- `-c <device_bus_id>` or `--ccw <device_bus_id>`
  filters LUNs by one or more adapters with the specified FCP device-bus IDs. When used in conjunction with `-p`, only those LUNs are listed that also satisfy at least one of the `-p` constraints.

- `-p <wwpn>` or `--port <wwpn>`
  filters LUNs by one or more target ports with the specified WWPNs. When used in conjunction with `-c`, only those LUNs are listed that also satisfy at least one of the `-c` constraints.

- `-a` or `--active`
  shows the encryption state of the attached LUNs. Encrypted devices are indicated with a bracketed X immediately after the LUN number.

- `-v` or `--version`
  displays version information and exits.

- `-h` or `--help`
  displays an overview of the syntax. To view the man page, enter `man islns`.

Examples

- This example lists all LUNs discovered in the FC SAN on adapter 0.0.3922:
  ```
  # islns -c 0.0.3922
  ```

- This example shows all LUNs discovered in the FC SAN on target port 0x500507630300c562:
  ```
  # islns -p 0x500507630300c562
  Scanning for LUNs on adapter 0.0.5922 at port 0x500507630300c562:
  0x4010400000000000
  0x4010400100000000
  0x4010400200000000
  0x4010400300000000
  0x4010400400000000
  0x4010400500000000
  ```

- This example shows all LUNs discovered in the FC SAN on:
  ```
  - Adapter 0.0.3922 and port 0x5005123456789000
  - Adapter 0.0.3922 and port 0x5005abcdefabc000
  - Adapter 0.0.fc00 and port 0x5005123456789000
  - Adapter 0.0.fc00 and port 0x5005abcdefabc000
  ```
This example shows the encryption status of attached LUNs:

```
# lsluns -a
adapter = 0.0.3c02
 port = 0x500507630303c562
 lun = 0x401040a200000000(X) /dev/sg0 Disk IBM:2107900
 lun = 0x401040a300000000 /dev/sg1 Disk IBM:2107900
...
 port = 0x500507630303c562
...
adapter = 0.0.593a
...
```

The (X) after the LUN number indicates that the device is encrypted.
Ismem - Show online status information about memory blocks

Use the `lsmem` command to list the ranges of available memory with their online status.

The listed memory blocks correspond to the memory block representation in sysfs. The command also shows the memory block size, the device size, and the amount of memory in online and offline state.

**Ismem syntax**

```
{lsmem} [-a]  
```

Where:

- `-a` or `--all` lists each individual memory block, instead of combining memory blocks with similar attributes.
- `-v` or `--version` displays the version number of `lsmem`, then exits.
- `-h` or `--help` displays a short help text, then exits. To view the man page, enter `man lsmem`.

The columns in the command output have this meaning:

**Address range**
Start and end address of the memory range.

**Size**
Size of the memory range in MB (1024 x 1024 bytes).

**State**
Indication of the online status of the memory range. State `on->off` means that the address range is in transition from online to offline.

**Removable**
Yes if the memory range can be set offline, `no` if it cannot be set offline. A dash (`-`) means that the range is already offline. The kernel method that identifies removable memory ranges is heuristic and not exact. Occasionally, memory ranges are falsely reported as removable or falsely reported as not removable.

**Device**
Device number or numbers that correspond to the memory range.

A device represents a unit of memory for the hypervisor in control of the memory. The hypervisor cannot reuse a device unless the entire corresponding memory range is offline.

The memory units that you can set online or offline from Linux are memory blocks. In most memory configurations, there is a one-to-one mapping of devices and memory blocks or a mapping of multiple devices to a single memory block. In other configurations, multiple memory blocks might map to a single device. Memory might be used inefficiently if a device includes both online and offline memory blocks.
The `chmem` command with the size parameter automatically chooses the best suited device or devices for setting memory online or offline. The device size depends on the hypervisor and on the amount of total online and offline memory.

**Examples**
- The output of this command, shows ranges of adjacent memory blocks with similar attributes.

```
# lsmem
Address range Size (MB) State Removable Device
0x0000000000000000-0x000000000fffffff 256 online no 0
0x0000000010000000-0x000000002fffffff 512 online yes 1-2
0x0000000030000000-0x000000003fffffff 256 online no 3
0x0000000040000000-0x000000006fffffff 768 online yes 4-6
0x0000000070000000-0x00000000fffffff 2304 offline - 7-15
```

Memory device size: 256 MB
Memory block size: 256 MB
Total online memory: 1792 MB
Total offline memory: 2304 MB

- The output of this command, shows each memory block as a separate range.

```
# lsmem -a
Address range Size (MB) State Removable Device
0x0000000000000000-0x000000000fffffff 256 online no 0
0x0000000010000000-0x000000001fffffff 256 online yes 1
0x0000000020000000-0x000000003fffffff 256 online yes 2
0x0000000030000000-0x000000003fffffff 256 online no 3
0x0000000040000000-0x000000004fffffff 256 online yes 4
0x0000000050000000-0x000000005fffffff 256 online yes 5
0x0000000060000000-0x000000006fffffff 256 online yes 6
0x0000000070000000-0x000000007fffffff 256 offline - 7
0x0000000080000000-0x000000008fffffff 256 offline - 8
0x0000000090000000-0x000000009fffffff 256 offline - 9
0x00000000a0000000-0x00000000afffffff 256 offline - 10
0x00000000b0000000-0x00000000bfffffffff 256 offline - 11
0x00000000c0000000-0x00000000cfffffff 256 offline - 12
0x00000000d0000000-0x00000000fffffff 256 offline - 13
0x00000000e0000000-0x00000000efffffffff 256 offline - 14
0x00000000f0000000-0x00000000fffffff 256 offline - 15
```

Memory device size: 256 MB
Memory block size: 256 MB
Total online memory: 1792 MB
Total offline memory: 2304 MB
Isqeth - List qeth-based network devices

Use the lsqeth command to display a summary of information about qeth-based network devices.

**Before you begin:** To be able to use this command, you must also install qethconf (see “qethconf - Configure qeth devices” on page 610). You install qethconf and lsqeth with the s390utils RPM.

Isqeth syntax

```
lsqeth [-p] [ <interface> ]
```

Where:

- **-p or --proc**
  displays the interface information in the former /proc/qeth format. This option can generate input to tools that expect this particular format.

- **<interface>**
  limits the output to information about the specified interface only.

- **-h or --help**
  displays a short help text, then exits.

**Examples**

- The following command lists information about interface enccw0.0.f5a0 in the default format:

  ```
  # lsqeth enccw0.0.f5a0
  Device name : enccw0.0.f5a0
  ---------------------------------------------
  card_type : 050_1000
  cdev0 : 0.0.f5a0
  cdev1 : 0.0.f5a3
  cdev2 : 0.0.f5a4
  chpid : B5
  online : 1
  portname : OSAPORT
  portno : 0
  route4 : no
  route6 : no
  checksumming : hw checksumming
  state : UP (LAN ONLINE)
  priority_queueing : always queue 2
  fake_broadcast : 0
  buffer_count : 64
  layer2 : 0
  large_send : no
  isolation : none
  sniffer : 0
  ```

- The following command lists information about all qeth-based interfaces in the former /proc/qeth format:
<table>
<thead>
<tr>
<th>devices</th>
<th>CHPID</th>
<th>interface</th>
<th>cardtype</th>
<th>port</th>
<th>chksum</th>
<th>prio-q'ing</th>
<th>rtr4</th>
<th>rtr6</th>
<th>lay'2</th>
<th>cnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.833c/0.0.8340/0.0.8341</td>
<td>xF8</td>
<td>encw0.0.833c</td>
<td>HiperSockets</td>
<td>0</td>
<td>sw</td>
<td>always_q_2</td>
<td>no</td>
<td>no</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>0.0.f5a0/0.0.f5a3/0.0.f5a4</td>
<td>xB5</td>
<td>encw0.0.f5a0</td>
<td>OSD_1000</td>
<td>0</td>
<td>hw</td>
<td>always_q_2</td>
<td>no</td>
<td>no</td>
<td>0</td>
<td>64</td>
</tr>
<tr>
<td>0.0.fba2/0.0.fba3/0.0.fba4</td>
<td>xB0</td>
<td>encw0.0.fba2</td>
<td>050_1000</td>
<td>0</td>
<td>sw</td>
<td>always_q_2</td>
<td>no</td>
<td>no</td>
<td>1</td>
<td>64</td>
</tr>
</tbody>
</table>
lsreipl - List IPL and re-IPL settings

Use the `lsreipl` command to find out which boot device and which options are used if you issue the reboot command.

You can also display information about the current boot device.

**lsreipl syntax**

```
lsreipl [-v] -i
```

Where:
- **-i or --ipl** displays the IPL setting.
- **-v or --version** displays the version number of `lsreipl` and exits.
- **-h or --help** displays an overview of the syntax. Any other parameters are ignored.

By default the re-IPL device is set to the current IPL device. Use the `chreipl` command to change the re-IPL settings.

**Examples**

- This example shows the current re-IPL settings:

  ```
  # lsreipl
  Re-IPL type: fcp
  WWPN: 0x500507630300c562
  LUN: 0x401040b300000000
  Device: 0.0.1700
  bootprog: 0
  br_lba: 0
  Bootparms: 
  ```
**Isscm - List storage-class memory increments**

Use the **Isscm** command to list status and other information about available storage-class memory increments.

**Isscm syntax**

```
--isscm
  -h
  -v
```

Where:

- **-h or --help**
  displays help information for the command. To view the man page, enter `man isscm`.

- **-v or --version**
  displays version information for the command.

In the output table, the columns have the following meaning:

- **SCM Increment**
  Starting address of the storage-class memory increment.

- **Size**
  Size of the block device that represents the storage-class memory increment.

- **Name**
  Name of the block device that represents the storage-class memory increment.

- **Rank**
  A quality ranking in the form of a number in the range 1 - 15 where a lower number means better ranking.

- **D_state**
  Data state of the storage-class memory increment. A number that indicates whether there is data on the increment. The data state can be:

  1. The increment contains zeros only.
  2. Data was written to the increment.
  3. No data was written to the increment since the increment was attached.

- **O_state**
  Operation state of the storage-class memory increment.

- **Pers**
  Persistence attribute.

- **ResID**
  Resource identifier.
Examples

- This command lists all increments:

```
# lsscm
SCM Increment Size Name Rank D_state O_state Pers ResID
0000000000000000 16384MB scma 1 2 1 2 1
0000000400000000 16384MB scmb 1 2 1 2 1
```
Isshut - List the current system shutdown actions

Use the `lsshut` command to see how the Linux instance is configured for the `halt`, `poff`, `reboot`, `restart`, and `panic` system shutdown triggers.

For more information about the shutdown triggers and possible shutdown actions, see Chapter 8, “Shutdown actions,” on page 109.

If the action is `kdump`, a second action might be listed. This second action is the backup action that is taken if `kdump` fails. See Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711 for details about using `kdump`.

Isshut syntax

```
  lsshut [-h] [-v]
```

Where:

- `-h` or `--help` displays a short help text, then exits.
- `-v` or `--version` displays the version number of `lsshut` and exits.

Examples

- To query the configuration issue:

  ```
  # lsshut
  Trigger          Action
  ================
  Halt             stop
  Power off        vmcmd (LOGOFF)
  Reboot           reipl
  Restart          kdump,dump_reipl
  Panic            kdump,dump_reipl
  ```
Istape - List tape devices

Use the `istape` command to gather information about tape devices and display it in a summary format.

It gathers information about CCW-attached tape devices and tape devices that are attached to the SCSI bus from sysfs (see "Displaying tape information" on page 210).

For information about SCSI tape devices, the command uses the following sources for the information displayed:

- The IBMtape or the open source lin_tape driver.
- The sg_inq command from the scsi/sg3_utils package.
- The st (SCSI tape) device driver in the Linux kernel.

If you use the IBMtape or lin_tape driver, the sg_inq utility is required. If sg_inq is missing, certain information about the IBMtape or lin_tape driver cannot be displayed.

Istape syntax

```
-istape -s, -t<devicetype>, --online, --offline

-<device_bus_id>, --ccw-only, --scsi-only, --verbose

Notes:
1 specify the first device bus-ID with a leading blank.
```

Where:

- `-s` or `--shortid`
  strips the “0.<n>.” from the device bus-IDs in the command output. For CCW-attached devices only.

- `-t` or `--type`
  limits the output to information about the specified type or types of CCW-attached devices only.

- `--ccw-only`
  limits the output to information about CCW-attached devices only.

- `--scsi-only`
  limits the output to information about tape devices that are attached to the SCSI bus.
lstape

`--online | --offline`
limits the output to information about online or offline CCW-attached tape devices only.

`<device_bus_id>`
limits the output to information about the specified tape device or devices only.

`--verbose`
For tape devices attached to the SCSI bus only. Prints the serial of the tape and information about the FCP connection as an additional text line after each SCSI tape in the list.

`-h or --help`
displays a short help text.

`--version`
displays the version of the command.

Examples

- This command displays information about all tapes that are found, here one CCW-attached tape and one tape and changer device that is configured for zFCP:

```
#> lstape
FICON/ESCON tapes (found 1):
TapeNo BusID CuType/Model DevType/Model BlkSize State Op MedState
 0 0.0.0480 3480/01 3480/04 auto UNUSED --- UNLOADED

SCSI tape devices (found 2):
Generic Device Target Vendor Model Type State
sg4 IBMchanger0 0:0:0:0 IBM 03590H11 changer running
sg5 IBMtape0 0:0:0:1 IBM 03590H11 tapedrv running
```

If only the generic tape driver (st) and the generic changer driver (ch) are loaded, the output lists those names in the device section:

```
#> lstape
FICON/ESCON tapes (found 1):
TapeNo BusID CuType/Model DevType/Model BlkSize State Op MedState
 0 0.0.0480 3480/01 3480/04 auto UNUSED --- UNLOADED

SCSI tape devices (found 2):
Generic Device Target Vendor Model Type State
sg0 sch0 0:0:0:0 IBM 03590H11 changer running
sg1 st0 0:0:0:1 IBM 03590H11 tapedrv running
```

- This command displays information about all available CCW-attached tapes.

```
#> lstape --ccw-only
TapeNo BusID CuType/Model DevType/DevMod BlkSize State Op MedState
 0 0.0.0132 3590/50 3590/11 auto IN_USE --- LOADED
 1 0.0.0110 3490/10 3490/40 auto UNUSED --- UNLOADED
 2 0.0.0133 3590/50 3590/11 auto IN_USE --- LOADED
 3 0.0.012a 3480/01 3480/04 auto UNUSED --- UNLOADED
N/A 0.0.01f8 3480/01 3480/04 N/A OFFLINE --- N/A
```

- This command limits the output to tapes of type 3480 and 3490.
# lstape

<table>
<thead>
<tr>
<th>TapeNo</th>
<th>BusID</th>
<th>CuType/Model</th>
<th>DevType/DevMod</th>
<th>BlkSize</th>
<th>State</th>
<th>Op</th>
<th>MedState</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0.0110</td>
<td>3490/10</td>
<td>3490/40</td>
<td>auto</td>
<td>UNUSED</td>
<td>---</td>
<td>UNLOADED</td>
</tr>
<tr>
<td>3</td>
<td>0.0.012a</td>
<td>3480/01</td>
<td>3480/04</td>
<td>auto</td>
<td>UNUSED</td>
<td>---</td>
<td>UNLOADED</td>
</tr>
<tr>
<td>N/A</td>
<td>0.0.01f8</td>
<td>3480/01</td>
<td>3480/04</td>
<td>N/A</td>
<td>OFFLINE</td>
<td>---</td>
<td>N/A</td>
</tr>
</tbody>
</table>

- This command limits the output to those tapes of type 3480 and 3490 that are currently online.

# lstape -t 3480,3490 --online

<table>
<thead>
<tr>
<th>TapeNo</th>
<th>BusID</th>
<th>CuType/Model</th>
<th>DevType/DevMod</th>
<th>BlkSize</th>
<th>State</th>
<th>Op</th>
<th>MedState</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0.0110</td>
<td>3490/10</td>
<td>3490/40</td>
<td>auto</td>
<td>UNUSED</td>
<td>---</td>
<td>UNLOADED</td>
</tr>
<tr>
<td>3</td>
<td>0.0.012a</td>
<td>3480/01</td>
<td>3480/04</td>
<td>auto</td>
<td>UNUSED</td>
<td>---</td>
<td>UNLOADED</td>
</tr>
</tbody>
</table>

- This command limits the output to the tape with device bus-ID 0.0.012a and strips the “0.<n>.” from the device bus-ID in the output.

# lstape -s 0.0.012a

<table>
<thead>
<tr>
<th>TapeNo</th>
<th>BusID</th>
<th>CuType/Model</th>
<th>DevType/DevMod</th>
<th>BlkSize</th>
<th>State</th>
<th>Op</th>
<th>MedState</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>012a</td>
<td>3480/01</td>
<td>3480/04</td>
<td>auto</td>
<td>UNUSED</td>
<td>---</td>
<td>UNLOADED</td>
</tr>
</tbody>
</table>

- This command limits the output to SCSI devices but gives more details. The serial numbers are only displayed if the `sg_inq` command is found on the system.

# lstape --scsi-only --verbose

<table>
<thead>
<tr>
<th>Generic Device</th>
<th>Target</th>
<th>Vendor</th>
<th>Model</th>
<th>Type</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBA: sg0</td>
<td>st0</td>
<td>IBM</td>
<td>03590H11</td>
<td>tapedrv</td>
<td>running</td>
</tr>
<tr>
<td>WWPN: 0.0.1708</td>
<td>0x500507630040727b</td>
<td>NO/INQ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWPN: sg1</td>
<td>sch0</td>
<td>IBM</td>
<td>03590H11</td>
<td>changer</td>
<td>running</td>
</tr>
<tr>
<td>WWPN: 0.0.1708</td>
<td>0x500507630040727b</td>
<td>NO/INQ</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data fields for SCSI tape devices**

There are specific data fields for SCSI tape devices.

**Table 61. lstape data fields for SCSI tape devices**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>SCSI generic device file for the tape drive (for example /dev/sg0). This attribute is empty if the <code>sg_inq</code> command is not available.</td>
</tr>
<tr>
<td>Device</td>
<td>Main device file for accessing the tape drive, for example:</td>
</tr>
<tr>
<td></td>
<td>- /dev/st0 for a tape drive that is attached through the Linux st device driver</td>
</tr>
<tr>
<td></td>
<td>- /dev/sch0 for a medium changer device that is attached through the Linux changer device driver</td>
</tr>
<tr>
<td></td>
<td>- /dev/IBMchanger0 for a medium changer that is attached through the IBMtape or lin_tape device driver</td>
</tr>
<tr>
<td></td>
<td>- /dev/IBMtape0 for a tape drive that is attached through the IBMtape or lin_tape device driver</td>
</tr>
<tr>
<td>Target</td>
<td>The ID in Linux used to identify the SCSI device.</td>
</tr>
<tr>
<td>Vendor</td>
<td>The vendor field from the tape drive.</td>
</tr>
<tr>
<td>Model</td>
<td>The model field from the tape drive.</td>
</tr>
<tr>
<td>Type</td>
<td>&quot;Tapedrv&quot; for a tape driver or &quot;changer&quot; for a medium changer.</td>
</tr>
<tr>
<td>State</td>
<td>The state of the SCSI device in Linux. This state is an internal state of the Linux kernel, any state other than &quot;running&quot; can indicate problems.</td>
</tr>
</tbody>
</table>
### Table 61. Iastape data fields for SCSI tape devices (continued)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBA</td>
<td>The FCP device to which the tape drive is attached.</td>
</tr>
<tr>
<td>WWPN</td>
<td>The WWPN (worldwide port name) of the tape drive in the SAN.</td>
</tr>
<tr>
<td>Serial</td>
<td>The serial number field from the tape drive.</td>
</tr>
</tbody>
</table>
lszcrypt - Display cryptographic devices

Use the lszcrypt command to display information about cryptographic adapters that are managed by the cryptographic device driver and its AP bus attributes.

To set the attributes, use “chzcrypt - Modify the cryptographic configuration” on page 503. The following information can be displayed for each cryptographic adapter:
- The card type
- The online status
- The hardware card type
- The card capability
- The hardware queue depth
- The request count

The following AP bus attributes can be displayed:
- The default AP domain
- The configuration timer
- The poll thread status
- The poll timeout
- The AP interrupt status

lszcrypt syntax

```
lszcrypt
  -b
  -c <device_ID>
  -d
  -V <device_ID>
```

Where:

- `<device ID>` specifies a cryptographic adapter to display. A cryptographic device can be either an adapter ID or an AP queue device. If no devices are specified, information about all available devices is displayed. Both the adapter ID representation and the AP queue device representation are hexadecimal.

- `-b` or `--bus` displays the AP bus attributes.

- `-c <device ID>` or `--capability <device ID>` shows the capabilities of a cryptographic adapter as of Crypto Express2 (CEX2). The capabilities of a cryptographic adapter depend on the card type and the installed function facilities. A cryptographic adapter can provide one or more of the following capabilities:
  - RSA 2K Clear Key
  - RSA 4K Clear Key
lszcrypt

- CCA Secure Key
- Long RNG

\textbf{-d} or \textbf{--domains}

shows the usage and control domains of the cryptographic device. The displayed domains of the cryptographic device depends on the initial cryptographic configuration.

- “C” indicates a control domain
- “U” indicates a usage domain
- “B” indicates both (control and usage domain)

\textbf{-V} or \textbf{--verbose}

enables the verbose level for cryptographic device information. It displays card type, online status, hardware card type, hardware queue depth, request count, pending request queue count, outstanding request queue count, and installed function facilities.

\textbf{-h} or \textbf{--help}

displays short information about command usage.

\textbf{-v} or \textbf{--version}

displays version information.

\section*{Examples}

These examples illustrate common uses for \texttt{lszcrypt}.

- To display information about all available cryptographic devices and AP queues:

\begin{verbatim}
# lszcrypt
\end{verbatim}

This command lists all devices grouped by cryptographic device, similar to the following example. The domain IDs are hexadecimal values.

\begin{center}
\begin{tabular}{lllll}
\textbf{CARD\_DOMAIN} & \textbf{TYPE} & \textbf{MODE} & \textbf{STATUS} & \textbf{REQUEST\_CNT} \\
\hline
00 & CEX5A & Accelerator & online & 0 \\
00.0005 & CEX5A & Accelerator & online & 0 \\
00.001f & CEX5A & Accelerator & online & 0 \\
00.004d & CEX5A & Accelerator & online & 0 \\
01 & CEX5C & CCA-Coproc & online & 26 \\
01.0005 & CEX5C & CCA-Coproc & online & 10 \\
01.001f & CEX5C & CCA-Coproc & online & 7 \\
01.004d & CEX5C & CCA-Coproc & online & 9 \\
05 & CEX5P & EP11-Coproc & online & 0 \\
05.0005 & CEX5P & EP11-Coproc & online & 0 \\
05.001f & CEX5P & EP11-Coproc & online & 0 \\
05.004d & CEX5P & EP11-Coproc & online & 0 \\
\hline
\end{tabular}
\end{center}

- To display AP bus information:

\begin{verbatim}
# lszcrypt -b
\end{verbatim}

This command displays output similar to the following example:

\begin{verbatim}
ap_domain=0x6
ap_max_domain_id=0x54
ap_interruption are enabled
config_time=30 (seconds)
poll_thread is disabled
poll_timeout=250000 (nanoseconds)
\end{verbatim}
To display the capabilities for the cryptographic device with adapter ID 0x0b:

```
# lszcrypt -c 0x0b
```

This command displays output similar to the following example:

```
Coprocessor card0b provides capability for:
CCA Secure Key
RSA 4K Clear Key
Long RNG
```

To list the usage and control domains of the cryptographic devices:

```
# lszcrypt -d
```

This command displays a table that lists all domains (in hex notation) similar to the following example:

```
<table>
<thead>
<tr>
<th>DOMAIN</th>
<th>00</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
<th>0a</th>
<th>0b</th>
<th>0c</th>
<th>0d</th>
<th>0e</th>
<th>0f</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>00</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>a0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>b0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>c0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>d0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>e0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>f0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
```

C: Control domain
U: Usage domain
B: Both (Control + Usage domain)

To display detailed information of all available cryptographic devices:

```
# lszcrypt -V
```

This example shows CEX5S cryptographic devices in accelerator mode, CCA and EP1 in coprocessor mode (IDs 0x00, 0x01 and 0x05) with domains 5 (0x05), 31 (0x1f) and 77 (0x4d) configured. Adapter IDs and domain IDs are hexadecimal values.
### lszcrypt

<table>
<thead>
<tr>
<th>CARD.DOMAIN</th>
<th>TYPE</th>
<th>MODE</th>
<th>STATUS</th>
<th>REQUEST_CNT</th>
<th>PENDINGQ_CNT</th>
<th>REQUESTQ_CNT</th>
<th>HW_TYPE</th>
<th>Q_DEPTH</th>
<th>FUNCTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>CEX5A</td>
<td>Accelerator</td>
<td>online</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x6a000000</td>
</tr>
<tr>
<td>00.0005</td>
<td>CEX5A</td>
<td>Accelerator</td>
<td>online</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x6a000000</td>
</tr>
<tr>
<td>00.001f</td>
<td>CEX5A</td>
<td>Accelerator</td>
<td>online</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x6a000000</td>
</tr>
<tr>
<td>00.004d</td>
<td>CEX5A</td>
<td>Accelerator</td>
<td>online</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x6a000000</td>
</tr>
<tr>
<td>01</td>
<td>CEXSC</td>
<td>CCA-Coproc</td>
<td>online</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x92000000</td>
</tr>
<tr>
<td>01.0005</td>
<td>CEXSC</td>
<td>CCA-Coproc</td>
<td>online</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x92000000</td>
</tr>
<tr>
<td>01.001f</td>
<td>CEXSC</td>
<td>CCA-Coproc</td>
<td>online</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x92000000</td>
</tr>
<tr>
<td>01.004d</td>
<td>CEXSC</td>
<td>CCA-Coproc</td>
<td>online</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x92000000</td>
</tr>
<tr>
<td>03</td>
<td>CEXSP</td>
<td>EP11-Coproc</td>
<td>online</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x06000000</td>
</tr>
<tr>
<td>03.0005</td>
<td>CEXSP</td>
<td>EP11-Coproc</td>
<td>online</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x06000000</td>
</tr>
<tr>
<td>03.001f</td>
<td>CEXSP</td>
<td>EP11-Coproc</td>
<td>online</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x06000000</td>
</tr>
<tr>
<td>03.004d</td>
<td>CEXSP</td>
<td>EP11-Coproc</td>
<td>online</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x06000000</td>
</tr>
</tbody>
</table>

To display the device ID and the installed function facility in hexadecimal notation as well as card type, online status, hardware card type, hardware queue depth, request count, pending request queue count, outstanding request queue count, and installed function facilities:

```sh
# lszcrypt -V 0x00 0x03 0x0b
```

This command displays output similar to the following example:

<table>
<thead>
<tr>
<th>CARD.DOMAIN</th>
<th>TYPE</th>
<th>MODE</th>
<th>STATUS</th>
<th>REQUEST_CNT</th>
<th>PENDINGQ_CNT</th>
<th>REQUESTQ_CNT</th>
<th>HW_TYPE</th>
<th>Q_DEPTH</th>
<th>FUNCTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>CEX5A</td>
<td>Accelerator</td>
<td>online</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x6a000000</td>
</tr>
<tr>
<td>00.0041</td>
<td>CEX5A</td>
<td>Accelerator</td>
<td>online</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x6a000000</td>
</tr>
<tr>
<td>03</td>
<td>CEXSC</td>
<td>CCA-Coproc</td>
<td>online</td>
<td>1631</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x92000000</td>
</tr>
<tr>
<td>03.0041</td>
<td>CEXSC</td>
<td>CCA-Coproc</td>
<td>online</td>
<td>1631</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x92000000</td>
</tr>
<tr>
<td>0b</td>
<td>CEXSP</td>
<td>EP11-Coproc</td>
<td>online</td>
<td>3088</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x06000000</td>
</tr>
<tr>
<td>0b.0041</td>
<td>CEXSP</td>
<td>EP11-Coproc</td>
<td>online</td>
<td>3088</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>08</td>
<td>0x06000000</td>
</tr>
</tbody>
</table>

**Tip:** In the device specification you can also use one-digit hexadecimal or decimal notation. The following specifications are all equivalent:

- 0x0 0x2 0xb
- 0x00 0x02 0x0b
- 0 2 11
lszfcp - List zfcp devices

Use the `lszfcp` command to gather information about zfcp devices, ports, units, and their associated class devices from sysfs and to display it in a summary format.

**lszfcp syntax**

```
lszfcp [options]
```

Where:

- `-H` or `--hosts`
  - shows information about hosts.

- `-P` or `--ports`
  - shows information about ports.

- `-D` or `--devices`
  - shows information about SCSI devices.

- `-a` or `--attributes`
  - shows all attributes (implies `-V`).

- `-V` or `--verbose`
  - shows sysfs paths of associated class and bus devices.

- `-b device_bus_id` or `--busid device_bus_id`
  - limits the output to information about the specified device.

- `-p port_name` or `--wwpn port_name`
  - limits the output to information about the specified port name.

- `-l lun` or `--lun lun`
  - limits the output to information about the specified LUN.

- `-s mount_point` or `--sysfs mount_point`
  - specifies the mount point for sysfs.

- `-v` or `--version`
  - displays version information.

- `-h` or `--help`
  - displays a short help text.
Examples

- This command displays information about all available hosts, ports, and SCSI devices.

```bash
# lszfcp -H -D -P
0.0.3d0c host0
0.0.500c host1
... 
0.0.3c0c host5
0.0.3d0c/0x500507630300c562 rport-0:0-0
0.0.3d0c/0x50050763030bc562 rport-0:0-1
0.0.3d0c/0x500507630303c562 rport-0:0-2
0.0.500c/0x50050763030bc562 rport-1:0-0
... 
0.0.3c0c/0x500507630303c562 rport-5:0-2
0.0.3d0c/0x500507630300c562/0x4010403200000000 0:0:0:0
0.0.3d0c/0x50050763030bc562/0x4010403200000000 0:0:1:0
0.0.3d0c/0x500507630303c562/0x4010403200000000 0:0:2:0
0.0.500c/0x50050763030bc562/0x4010403200000000 1:0:0:0
... 
0.0.3c0c/0x500507630303c562/0x4010403200000000 5:0:2:0
```

- This command shows SCSI devices and limits the output to the devices that are attached through the FCP device with bus ID 0.0.3d0c:

```bash
# lszfcp -D -b 0.0.3d0c
0.0.3d0c/0x500507630300c562/0x4010403200000000 0:0:0:0
0.0.3d0c/0x50050763030bc562/0x4010403200000000 0:0:0:1
0.0.3d0c/0x500507630303c562/0x4010403200000000 0:0:0:2
0.0.3d0c/0x50050763030bc562/0x4010403200000000 1:0:0:0
```
The `mon_fsstatd` command is a user space daemon that collects physical file system size data from Linux on z/VM.

The daemon periodically writes the data as defined records to the z/VM monitor stream using the monwriter character device driver. You can start the daemon with a service script `/etc/init.d/mon_statd` or call it manually. When it is called with the service utility, it reads the configuration file `/etc/sysconfig/mon_statd`.

**Before you begin:**

- Install the monwriter device driver and set up z/VM to start the collection of monitor sample data. See Chapter 31, “Writing z/VM monitor records,” on page 381 for information about the setup for and usage of the monwriter device driver.
- Customize the configuration file `/etc/sysconfig/mon_statd` if you plan to call it with the service utility.

The following books provide general information about DCSSs, DIAG x'DC', CP commands, and APPLDATA:

- See z/VM Saved Segments Planning and Administration, SC24-6229 for general information about DCSSs.
- See z/VM CP Programming Services, SC24-6179 for information about the DIAG x'DC' instruction.
- See z/VM CP Commands and Utilities Reference, SC24-6175 for information about the CP commands.
- See z/VM Performance, SC24-6208 for information about monitor APPLDATA.

You can run the `mon_fsstatd` command in two ways:

- Calling `mon_statd` with the service utility. This method reads the configuration file `/etc/sysconfig/mon_statd`. The `mon_statd` service script also controls other daemons, such as `mon_procd`.
- Calling `mon_fsstatd` from a command line.

### mon_statd service utility syntax

If you run the `mon_fsstatd` daemon through the service utility, you configure the daemon through specifications in a configuration file.

```bash
service mon_statd
    [ /etc/init.d/mon_statd ]
        start
        stop
        status
        restart
```

Where:

**start**

enables monitoring of guest file system size, by using the configuration in `/etc/sysconfig/mon_statd`. 
**mon_fsstatd**

**stop**

Disables monitoring of guest file system size.

**status**

Shows current status of guest file system size monitoring.

**restart**

Stops and restarts monitoring. Useful to re-read the configuration file when it was changed.

**Configuration file keywords**

**FSSTAT_INTERVAL="<n>"**

Specifies the wanted sampling interval in seconds.

**FSSTAT="yes | no"**

Specifies whether to enable the mon_fsstatd daemon. Set to “yes” to enable the daemon. Anything other than “yes” is interpreted as “no”.

**Examples of service utility use**

Example configuration file for mon_statd (/etc/sysconfig/mon_statd).

- This example sets the sampling interval to 30 seconds and enables the mon_fsstatd daemon:

```
FSSTAT_INTERVAL="30"
FSSTAT="yes"
```

Example of mon_statd use. Note that your output can look different and include messages for other daemons, such as mon_procd:

- To enable guest file system size monitoring:

```
> service mon_statd start
... Starting mon_fsstatd: [ OK ]
...```

- To display the status:

```
> service mon_statd status
... mon_fsstatd (pid 1075, interval: 30) is running.
...```

- To disable guest file system size monitoring:

```
> service mon_statd stop
... Stopping mon_fsstatd: [ OK ]
...```

- To display the status again and check that monitoring is now disabled:

```
> service mon_statd status
... mon_fsstatd is not running
...```

- To restart the daemon and re-read the configuration file:
mon_fsstatd command-line syntax

If you call the mon_fsstatd daemon from the command line, you configure the daemon through command parameters.

Where:
- **-i or --interval <seconds>**
  specifies the wanted sampling interval in seconds.
- **-a or --attach**
  runs the daemon in the foreground.
- **-h or --help**
  displays help information for the command.
- **-v or --version**
  displays version information for the command.

**Examples of command-line use**

- To start mon_fsstatd with default setting:

  ```
  > mon_fsstatd
  ```

- To start mon_fsstatd with a sampling interval of 30 seconds:

  ```
  > mon_fsstatd -i 30
  ```

- To start mon_fsstatd and have it run in the foreground:

  ```
  > mon_fsstatd -a
  ```

- To start mon_fsstatd with a sampling interval of 45 seconds and have it run in the foreground:

  ```
  > mon_fsstatd -a -i 45
  ```

**Processing monitor data**

The mon_fsstatd daemon writes physical file system size data for Linux on z/VM to the z/VM monitor stream.
The following is the format of the file system size data that is passed to the z/VM monitor stream. One sample monitor record is written for each physical file system that is mounted at the time of the sample interval. The monitor data in each record contains a header consisting of a time stamp, the length of the data, and an offset. The header is followed by the file system data (as obtained from statvfs). The file system data fields begin with “fs_”.

**Table 62. File system size data format**

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>__u64</td>
<td>time_stamp</td>
<td>Time at which the file system data was sampled.</td>
</tr>
<tr>
<td>__u16</td>
<td>data_len</td>
<td>Length of data that follows the header.</td>
</tr>
<tr>
<td>__u16</td>
<td>data_offset</td>
<td>Offset from start of the header to start of file system data (that is, to the fields that begin with fs_).</td>
</tr>
<tr>
<td>__u16</td>
<td>fs_name_len</td>
<td>Length of the file system name. The file system name can be too long to fit in the monitor record. If so, this length is the portion of the name that is contained in the monitor record.</td>
</tr>
<tr>
<td>char [fs_name_len]</td>
<td>fs_name</td>
<td>The file system name. If the name is too long to fit in the monitor record, the name is truncated to the length in the fs_name_len field.</td>
</tr>
<tr>
<td>__u16</td>
<td>fs_dir_len</td>
<td>Length of the mount directory name. The mount directory name can be too long to fit in the monitor record. If so, this length is the portion of the name that is contained in the monitor record.</td>
</tr>
<tr>
<td>char [fs_dir_len]</td>
<td>fs_dir</td>
<td>The mount directory name. If the name is too long to fit in the monitor record, the name is truncated to the length in the fs_dir_len field.</td>
</tr>
<tr>
<td>__u16</td>
<td>fs_type_len</td>
<td>Length of the mount type. The mount type can be too long to fit in the monitor record. If so, this length is the portion that is contained in the monitor record.</td>
</tr>
<tr>
<td>char [fs_type_len]</td>
<td>fs_type</td>
<td>The mount type (as returned by getmntent). If the type is too long to fit in the monitor record, the type is truncated to the length in the fs_type_len field.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_bsize</td>
<td>File system block size.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_frsize</td>
<td>Fragment size.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_blocks</td>
<td>Total data blocks in file system.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_bfree</td>
<td>Free blocks in fs.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_bavail</td>
<td>Free blocks avail to non-superuser.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_files</td>
<td>Total file nodes in file system.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_ffree</td>
<td>Free file nodes in fs.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_favail</td>
<td>Free file nodes available to non-superuser.</td>
</tr>
<tr>
<td>__u64</td>
<td>fs_flag</td>
<td>Mount flags.</td>
</tr>
</tbody>
</table>

Use the time_stamp to correlate all file systems that were sampled in a given interval.
Reading the monitor data

All records that are written to the z/VM monitor stream begin with a product identifier.

The product ID is a 16-byte structure of the form `pppppppffnvrrmm`, where for records that are written by `mon_fsstatd`, these values are:

- `pppppppp` is a fixed ASCII string `LNXAPPL`.
- `ff` is the application number for `mon_fsstatd = x'0001`.
- `n` is the record number = `x'00'`.
- `vv` is the version number = `x'0000'`.
- `rr` is reserved for future use and should be ignored.
- `mm` is reserved for `mon_fsstatd` and should be ignored.

Note: Though the mod_level field (mm) of the product ID varies, there is no relationship between any particular mod_level and file system. The mod_level field should be ignored by the reader of this monitor data.

There are many tools available to read z/VM monitor data. One such tool is the Linux monreader character device driver. For more information about monreader, see Chapter 32, “Reading z/VM monitor records,” on page 385.

mon_procd – Monitor Linux on z/VM

The `mon_procd` command is a user space daemon that gathers system summary information and information about up to 100 concurrent processes on Linux on z/VM.

The daemon writes this data to the z/VM monitor stream using the monwriter character device driver. You can start the daemon with a service script `/etc/init.d/mon_statd` or call it manually. When it is called with the service utility, it reads the configuration file `/etc/sysconfig/mon_statd`.

Before you begin:

- Install the monwriter device driver and set up z/VM to start the collection of monitor sample data. See Chapter 31, “Writing z/VM monitor records,” on page 381 for information about the setup for and usage of the monwriter device driver.
- Customize the configuration file `/etc/sysconfig/mon_statd` if you plan to call it with the service utility.
- The Linux instance on which the proc_mond daemon runs requires a z/VM guest virtual machine with the OPTION APPLMON statement in the CP directory entry.

The following books provide general information about DCSSs, CP commands, and APPLDATA:

- See z/VM Saved Segments Planning and Administration, SC24-6229 for general information about DCSSs.
- See z/VM CP Commands and Utilities Reference, SC24-6175 for information about the CP commands.
- See z/VM Performance, SC24-6208 for information about monitor APPLDATA.
You can run the `mon_procd` command in two ways:

- Calling `mon_procd` with the service utility. Use this method when you have the `mon_statd` service script that is installed in `/etc/init.d`. This method reads the configuration file `/etc/sysconfig/mon_statd`. The `mon_statd` service script also controls other daemons, such as `mon_fsstatd`.
- Calling `mon_procd` manually from a command line.

### `mon_statd` service utility syntax

If you run the `mon_procd` daemon through the service utility, you configure the daemon through specifications in a configuration file.

```
  service mon_statd
  /etc/init.d/mon_statd
  start
  stop
  status
  restart
```

Where:

- **start** enables monitoring of guest process data, by using the configuration in `/etc/sysconfig/mon_statd`.
- **stop** disables monitoring of guest process data.
- **status** shows current status of guest process data monitoring.
- **restart** stops and restarts guest process data monitoring. Useful to re-read the configuration file after it changed.

### Configuration file keywords

- **PROC_INTERVAL="<n>"** specifies the wanted sampling interval in seconds.
- **PROC="yes | no"** specifies whether to enable the `mon_procd` daemon. Set to “yes” to enable the daemon. Anything other than “yes” is interpreted as “no”.

### Examples of service utility use

Example configuration file for `mon_statd` (`/etc/sysconfig/mon_statd`).

- This example sets the sampling interval to 30 seconds and enables the `mon_procd`:

````
  PROC_INTERVAL="30"
  PROC="yes"
```

Example of `mon_statd` use (note that your output might look different and include messages for other daemons, such as `mon_fsstatd`):
• To enable guest process data monitoring:

```
> service mon_statd start
...$Starting mon_procd:[ OK ]
...
```

• To display the status:

```
> service mon_statd status
...
mon_procd (pid 1075, interval: 30) is running.
...
```

• To disable guest process data monitoring:

```
> service mon_statd stop
...$Stopping mon_procd:[ OK ]
...
```

• To display the status again and check that monitoring is now disabled:

```
> service mon_statd status
...
mon_procd is not running
...
```

• To restart the daemon and re-read the configuration file:

```
> service mon_statd restart
...
stopping mon_procd:[ OK ]
starting mon_procd:[ OK ]
...
```

**mon_procd command-line syntax**

If you call the `mon_procd` daemon from the command line, you configure the daemon through command parameters.

```
mon_procd -i 60
```

Where:

- `-i <seconds>` or `--interval <seconds>`
  specifies the wanted sampling interval in seconds.

- `-a` or `--attach`
  runs the daemon in the foreground.

- `-h` or `--help`
  displays help information for the command.

- `-v` or `--version`
  displays version information for the command.
Examples of command-line use

- To start `mon_procd` with default setting:
  
  ```
  > mon_procd
  ```

- To start `mon_procd` with a sampling interval of 30 seconds:
  
  ```
  > mon_procd -i 30
  ```

- To start `mon_procd` and have it run in the foreground:
  
  ```
  > mon_procd -a
  ```

- To start `mon_procd` with a sampling interval of 45 seconds and have it run in the foreground:
  
  ```
  > mon_procd -a -i 45
  ```

Processing monitor data

The `mon_procd` daemon writes process data to the z/VM monitor stream.

The data includes system summary information and information of each process for up to 100 processes currently being managed by an instance of Linux on z/VM to the z/VM monitor stream. At the time of the sample interval, one sample monitor record is written for system summary data. Then, one sample monitor record is written for each process for up to 100 processes currently being managed by the Linux instance. If more than 100 processes exist in a Linux instance at a given time, processes are sorted by the sum of CPU and memory usage percentage values. Only the top 100 processes’ data is written to the z/VM monitor stream.

The monitor data in each record begins with a header (a time stamp, the length of the data, and the offset). The data after the header depends on the field ”record number” of the 16-bit product ID and can be summary data or process data. See “Reading the monitor data” on page 607 for details. The following is the format of system summary data that is passed to the z/VM monitor stream.

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>__u64</td>
<td>time_stamp</td>
<td>Time at which the process data was sampled.</td>
</tr>
<tr>
<td>__u16</td>
<td>data_len</td>
<td>Length of data that follows the header.</td>
</tr>
<tr>
<td>__u16</td>
<td>data_offset</td>
<td>Offset from start of the header to the start of the process data.</td>
</tr>
<tr>
<td>__u64</td>
<td>uptime</td>
<td>Uptime of the Linux instance.</td>
</tr>
<tr>
<td>__u32</td>
<td>users</td>
<td>Number of users on the Linux instance.</td>
</tr>
<tr>
<td>char[6]</td>
<td>loadavg_1</td>
<td>Load average over the last 1 minute.</td>
</tr>
<tr>
<td>char[6]</td>
<td>loadavg_5</td>
<td>Load average over the last 5 minutes.</td>
</tr>
<tr>
<td>char[6]</td>
<td>loadavg_15</td>
<td>Load average over the last 15 minutes.</td>
</tr>
<tr>
<td>__u32</td>
<td>task_total</td>
<td>total number of tasks on the Linux instance.</td>
</tr>
<tr>
<td>__u32</td>
<td>task_running</td>
<td>Number of running tasks.</td>
</tr>
<tr>
<td>__u32</td>
<td>task_sleeping</td>
<td>Number of sleeping tasks.</td>
</tr>
<tr>
<td>__u32</td>
<td>task_stopped</td>
<td>Number of stopped tasks.</td>
</tr>
</tbody>
</table>
### Table 63. System summary data format (continued)

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>__u32</td>
<td>task_zombie</td>
<td>Number of zombie tasks.</td>
</tr>
<tr>
<td>__u32</td>
<td>num_cpus</td>
<td>Number of CPUs.</td>
</tr>
<tr>
<td>__u16</td>
<td>puser</td>
<td>A number that represents ((100 \times \text{percentage of total CPU time used for normal processes executing in user mode})).</td>
</tr>
<tr>
<td>__u16</td>
<td>pnice</td>
<td>A number that represents ((100 \times \text{percentage of total CPU time used for niced processes executing in user mode})).</td>
</tr>
<tr>
<td>__u16</td>
<td>psystem</td>
<td>A number that represents ((100 \times \text{percentage of total CPU time used for processes executing in kernel mode})).</td>
</tr>
<tr>
<td>__u16</td>
<td>pidle</td>
<td>A number that represents ((100 \times \text{percentage of total CPU idle time})).</td>
</tr>
<tr>
<td>__u16</td>
<td>piowait</td>
<td>A number that represents ((100 \times \text{percentage of total CPU time used for I/O wait})).</td>
</tr>
<tr>
<td>__u16</td>
<td>pirq</td>
<td>A number that represents ((100 \times \text{percentage of total CPU time used for interrupts})).</td>
</tr>
<tr>
<td>__u16</td>
<td>psoftirq</td>
<td>A number that represents ((100 \times \text{percentage of total CPU time used for softirqs})).</td>
</tr>
<tr>
<td>__u16</td>
<td>psteal</td>
<td>A number that represents ((100 \times \text{percentage of total CPU time spent in stealing})).</td>
</tr>
<tr>
<td>__u64</td>
<td>mem_total</td>
<td>Total memory in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>mem_used</td>
<td>Used memory in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>mem_free</td>
<td>Free memory in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>mem_buffers</td>
<td>Memory in buffer cache in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>mem_pgpgin</td>
<td>Data that is read from disk in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>mem_pgpgout</td>
<td>Data that is written to disk in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>swap_total</td>
<td>Total swap memory in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>swap_used</td>
<td>Used swap memory in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>swap_free</td>
<td>Free swap memory in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>swap Cached</td>
<td>Cached swap memory in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>swap_pswpin</td>
<td>Pages that are swapped in.</td>
</tr>
<tr>
<td>__u64</td>
<td>swap_pswpout</td>
<td>Pages that are swapped out.</td>
</tr>
</tbody>
</table>

The following is the format of a process information data that is passed to the z/VM monitor stream.

### Table 64. Process data format

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>__u64</td>
<td>time_stamp</td>
<td>Time at which the process data was sampled.</td>
</tr>
<tr>
<td>__u16</td>
<td>data_len</td>
<td>Length of data following the header.</td>
</tr>
<tr>
<td>__u16</td>
<td>data_offset</td>
<td>Offset from start of the header to the start of the process data.</td>
</tr>
<tr>
<td>__u32</td>
<td>pid</td>
<td>ID of the process.</td>
</tr>
<tr>
<td>Type</td>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>__u32</td>
<td>ppid</td>
<td>ID of the process parent.</td>
</tr>
<tr>
<td>__u32</td>
<td>euid</td>
<td>Effective user ID of the process owner.</td>
</tr>
<tr>
<td>__u16</td>
<td>tty</td>
<td>Device number of the controlling terminal or 0.</td>
</tr>
<tr>
<td>__s16</td>
<td>priority</td>
<td>Priority of the process</td>
</tr>
<tr>
<td>__s16</td>
<td>nice</td>
<td>Nice value of the process.</td>
</tr>
<tr>
<td>__u32</td>
<td>processor</td>
<td>Last used processor.</td>
</tr>
<tr>
<td>__u16</td>
<td>pcpu</td>
<td>A number that represents (100 * percentage of the elapsed cpu time used by the process since last sampling).</td>
</tr>
<tr>
<td>__u16</td>
<td>pmem</td>
<td>A number that represents (100 * percentage of physical memory used by the process).</td>
</tr>
<tr>
<td>__u64</td>
<td>total_time</td>
<td>Total cpu time the process used.</td>
</tr>
<tr>
<td>__u64</td>
<td>ctotal_time</td>
<td>Total cpu time the process and its dead child processes used.</td>
</tr>
<tr>
<td>__u64</td>
<td>size</td>
<td>Total virtual memory that is used by the task in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>swap</td>
<td>Swapped out portion of the virtual memory in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>resident</td>
<td>Non-swapped physical memory that is used by the task in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>trs</td>
<td>Physical memory that is devoted to executable code in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>drs</td>
<td>Physical memory that is devoted to other than executable code in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>share</td>
<td>Shared memory that is used by the task in KB.</td>
</tr>
<tr>
<td>__u64</td>
<td>dt</td>
<td>Dirty page count.</td>
</tr>
<tr>
<td>__u64</td>
<td>maj_fl t</td>
<td>Number of major page faults occurred for the process.</td>
</tr>
<tr>
<td>char</td>
<td>state</td>
<td>Status of the process.</td>
</tr>
<tr>
<td>__u32</td>
<td>flags</td>
<td>The process current scheduling flags.</td>
</tr>
<tr>
<td>__u16</td>
<td>ruser_len</td>
<td>Length of real user name of the process owner and should not be larger than 64.</td>
</tr>
<tr>
<td>char[ruser_len]</td>
<td>ruser</td>
<td>Real user name of the process owner. If the name is longer than 64, the name is truncated to the length 64.</td>
</tr>
<tr>
<td>__u16</td>
<td>euser_len</td>
<td>Length of effective user name of the process owner and should not be larger than 64.</td>
</tr>
<tr>
<td>char[euser_len]</td>
<td>euser</td>
<td>Effective user name of the process owner. If the name is longer than 64, the name is truncated to the length 64.</td>
</tr>
<tr>
<td>__u16</td>
<td>egroup_len</td>
<td>Length of effective group name of the process owner and should not be larger than 64.</td>
</tr>
<tr>
<td>char[egroup_len]</td>
<td>egroup</td>
<td>Effective group name of the process owner. If the name is longer than 64, the name is truncated to the length 64.</td>
</tr>
<tr>
<td>__u16</td>
<td>wchan_len</td>
<td>Length of sleeping in function's name and should not be larger than 64.</td>
</tr>
<tr>
<td>char[wchan_len]</td>
<td>wchan_name</td>
<td>Name of sleeping in function or '-' If the name is longer than 64, the name is truncated to the length 64.</td>
</tr>
<tr>
<td>__u16</td>
<td>cmd_len</td>
<td>Length of command name or program name that is used to start the process and should not be larger than 64.</td>
</tr>
</tbody>
</table>
Table 64. Process data format (continued)

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>char[cmd_len]</td>
<td>cmd</td>
<td>Command or program name that is used to start the process. If the name is</td>
</tr>
<tr>
<td></td>
<td></td>
<td>longer than 64, the name is truncated to the length 64.</td>
</tr>
<tr>
<td>__u16</td>
<td>cmd_line_len</td>
<td>Length of command line that is used to start the process and should not be</td>
</tr>
<tr>
<td></td>
<td></td>
<td>larger than 1024.</td>
</tr>
<tr>
<td>char[cmd_line_len]</td>
<td>cmd_line</td>
<td>Command line that is used to start the process. If the name is longer than</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1024, the name is truncated to the length 1024.</td>
</tr>
</tbody>
</table>

Use the time_stamp to correlate all process information that were sampled in a given interval.

**Reading the monitor data**

All records written to the z/VM monitor stream begin with a product identifier.

The product ID is a 16-byte structure of the form ppppppppfnvrrmm, where for records that are written by mon_procd, these values are:

- **PPPPPPP**
  - is a fixed ASCII string LNXAPPL.
- **ff**
  - is the application number for mon_procd = x'0002'.
- **n**
  - is the record number as follows:
    - x'00' indicates summary data.
    - x'01' indicates process data.
- **vv**
  - is the version number = x'0000'.
- **rr**
  - is the release number, which can be used to mark different versions of process APPLDATA records.
- **mm**
  - is reserved for mon_procd and should be ignored.

**Note:** Though the mod_level field (mm) of the product ID varies, there is no relationship between any particular mod_level and process. The mod_level field should be ignored by the reader of this monitor data.

This item uses at most 101 monitor buffer records from the monwriter device driver. A maximum number of buffers is set when a monwriter module is loaded. Because of this, the maximum number of buffers must not be less than the sum of buffer records that are used by all monwriter applications.

There are many tools available to read z/VM monitor data. One such tool is the Linux monreader character device driver. For more information about monreader, see Chapter 32, “Reading z/VM monitor records,” on page 385.
Use the `qetharp` command to query and purge address data such as MAC and IP addresses from the ARP cache of the OSA and HiperSockets hardware.

Before you begin:
- The `qetharp` command applies only to devices in layer 3 mode (see "Layer 2 and layer 3" on page 229).
- The `qetharp` command supports IPv6 only for real HiperSockets and z/VM guest LAN HiperSockets.
- For HiperSockets, z/VM guest LAN and VSWITCH interfaces, the `qetharp` command supports only the `--query` option.

**qetharp syntax**

```
qetharp
  -q <interface>  
  -n  
  -6  
  -c  
  -a <interface>  
  -d <interface>  
  -p <interface>  
  -i <ip_address>  
  -m <mac_address>
```

Where:
- **-q or --query**  
  shows the address resolution protocol (ARP) information about the specified network interface. Depending on the device that the interface was assigned to, this information is obtained from an OSA feature's ARP cache or a HiperSockets ARP cache.

  The default command output shows symbolic host names and includes only numerical addresses for host names that cannot be resolved. Use the `-n` option to show numerical addresses instead of host names.

  By default, qetharp omits IPv6 related information. Use the `-6` option to include IPv6 information for HiperSockets.

- **-n or --numeric**  
  shows numeric addresses instead of trying to determine symbolic host names. This option can be used only with the `-q` option.

- **-c or --compact**  
  limits the output to numeric addresses only. This option can be used only with the `-q` option.

- **-6 or --ipv6**  
  includes IPv6 information for HiperSockets. For real HiperSockets, shows the IPv6 addresses. For guest LAN HiperSockets, shows the IPv6 to MAC address mappings. This option can be used only with the `-q` option.

- `<interface>`  
  specifies the qeth interface to which the command applies.

- **-a or --add**  
  adds a static ARP entry to the OSA adapter. Static entries can be deleted with `-d`. 


-d or --delete
deletes a static ARP entry from the OSA adapter. Static entries are created with
-a.

-p or --purge
flushes the ARP cache of the OSA. The cache contains dynamic ARP entries,
which the OSA adapter creates through ARP queries. After flushing the cache,
the OSA adapter creates new dynamic entries. This option works only with
OSA devices. qetharp returns immediately.

-i <ip_address> or --ip <ip_address>
specifies the IP address to be added to or removed from the OSA adapter.

-m <mac_address> or --mac <mac_address>
specifies the MAC address to be added to the OSA adapter.

-v or --version
shows version information and exits

-h or --help
displays usage information and exits. To view the man page, enter man
qetharp.

Examples
• Show all ARP entries of the OSA defined as enccw0.0.f500:

  # qetharp -q enccw0.0.f500

• Show all ARP entries of the HiperSockets interface that is defined as
  enccw0.0.a1c0 including IPv6 entries:

  qetharp -6q enccw0.0.a1c0

• Show all ARP entries of the OSA defined as enccw0.0.f500, without resolving
  host names:

  # qetharp -nq enccw0.0.f500

• Show all ARP entries, including IPv6 entries, of the HiperSockets interface that is
defined as enccw0.0.a1c0 without resolving host names:

  qetharp -n6q enccw0.0.a1c0

• Flush the OSA ARP cache for enccw0.0.f500:

  # qetharp -p enccw0.0.f500

• Add a static entry for enccw0.0.f500 and IP address 1.2.3.4 to the OSA ARP
cache, with MAC address aa:bb:cc:dd:ee:ff:

  # qetharp -a enccw0.0.f500 -i 1.2.3.4 -m aa:bb:cc:dd:ee:ff

• Delete the static entry for enccw0.0.f500 and IP address 1.2.3.4 from the OSA
ARP cache.

  # qetharp -d enccw0.0.f500 -i 1.2.3.4
qethconf - Configure qeth devices

Use the `qethconf` command to configure IP address takeover, virtual IP address (VIPA), and proxy ARP for layer3 qeth devices.

See Chapter 15, “qeth device driver for OSA-Express (QDIO) and HiperSockets,” on page 223 for details about the following concepts:

- IP address takeover
- VIPA (virtual IP address)
- Proxy ARP

You cannot use this command with the layer2 option.

From the arguments that are specified, `qethconf` assembles the function command and redirects it to the corresponding sysfs attributes. You can also use `qethconf` to list the already defined entries.

qethconf syntax

```
qethconf
```

The `qethconf` command has these function keywords:

- `ipa`
  - configures qeth for IP address takeover (IPA).

- `vipa`
  - configures qeth for virtual IP address (VIPA).

- `parp` or `rxip`
  - configures qeth for proxy ARP.

The `qethconf` command has these action keywords:

- `add`
  - adds an IP address or address range.

- `del`
  - deletes an IP address or address range.

- `inv4`
  - inverts the selection of address ranges for IPv4 address takeover. This inversion makes the list of IP addresses that was specified with `qethconf add` and `qethconf del` an exclusion list.
inv6
inverts the selection of address ranges for IPv6 address takeover. This inversion
makes the list of IP addresses that was specified with qethconf add and
qethconf del an exclusion list.

list
lists existing definitions for specified qeth function.

list_all
lists existing definitions for IPA, VIPA, and proxy ARP.

<ip_addr>
IP address. Can be specified in one of these formats:
• IP version 4 format, for example, 192.168.10.38
• IP version 6 format, for example, FE80::1:800:23e7:f5db
• 8- or 32-character hexadecimals prefixed with -x, for example, -xc0a80a26

<mask_bits>
specifies the number of bits that are set in the network mask. Enables you to
specify an address range.

Example: A <mask_bits> of 24 corresponds to a network mask of
255.255.255.0.

<intface>
specifies the name of the interface that is associated with the specified address
or address range.

list_msg
lists qethconf messages and explanations.

-h or --help
displays help information.

-v or --version
displays version information.

Examples
• List existing proxy ARP definitions:

```
# qethconf parp list
parp add 1.2.3.4 enccw0.0.f500
```

• Assume responsibility for packages that are destined for 1.2.3.5:

```
# qethconf parp add 1.2.3.5 enccw0.0.f500
gethconf: Added 1.2.3.5 to /sys/class/net/enccw0.0.f500/device/rxip/add4.
gethconf: Use "qethconf parp list" to check for the result
```

Confirm the new proxy ARP definitions:

```
# qethconf parp list
parp add 1.2.3.4 enccw0.0.f500
parp add 1.2.3.5 enccw0.0.f500
```

• Configure enccw0.0.f500 for IP address takeover for all addresses that start with
192.168.10:
Display the new IP address takeover definitions:

```
# qethconf ipa list
ipa add 192.168.10.0/24 enccw0.0.f500
```

- Configure VIPA for enccw0.0.e400:

```
# qethconf vipa add 10.99.3.3 enccw0.0.e400
qethconf: Added 10.99.3.3 to /sys/class/net/enccw0.0.e400/device/vipa/add4.
qethconf: Use "qethconf vipa list" to check for the result
```

Display the new VIPA definitions:

```
# qethconf vipa list
vipa add 10.99.3.3 enccw0.0.e400
```

- List all existing IPA, VIPA, and proxy ARP definitions.

```
# qethconf list_all
parp add 1.2.3.4 enccw0.0.f500
parp add 1.2.3.5 enccw0.0.f500
ipa add 192.168.10.0/24 enccw0.0.f500
vipa add 10.99.3.3 enccw0.0.e400
```
qethqoat - Query OSA address table

Purpose

Use the `qethqoat` command to query the OSA address table and display physical and logical device information.

qethqoat syntax

```
qethqoat <interface> -r -s 0 -h -v
```

where:

- `-r` or `--raw`
  writes raw data to stdout.

- `-s` or `--scope`
  defines the scope of the query. The following values are valid:

  0    queries the level of the OSA address table

  1    interface (this is the default)

- `<interface>`
  specifies the interface for which you want to display information.

- `-h` or `--help`
  displays help information. To view the man page, enter `man qethqoat`.

- `-v` or `--version`
  displays version information.

Examples

To display physical and logical device information for interface enccw0.0.f500, issue:
This example uses scope 0 to query the supported OAT level and descriptor header types.

```
# qethqoat enccw0.0.f500
PCHID: 0x0310
CHPID: 0xa9
Manufacturer MAC address: 6c:ae:8b:48:0b:68
Configured MAC address: 00:00:00:00:00:00
Data device sub-channel address: 0xf402
CULA: 0x00
Unit address: 0x02
Physical port number: 0
Number of output queues: 1
Number of input queues: 1
Number of active input queues: 0
CHPID Type: OSD
Interface flags: 0x0a000000
OSA Generation: OSA-Express5S
Port speed/mode: 10 Gb/s / full duplex
Port media type: single mode (LR/LX)
Jumbo frames: yes
Firmware: 0x00000c9a
IPv4 router: no
IPv6 router: no
IPv4 vmac router: no
IPv6 vmac router: no
Connection isolation: not active
Connection isolation VEPA: no
IPv4 assists enabled: 0x00011c77
IPv6 assists enabled: 0x00f15c60
IPv4 outbound checksum enabled: 0x0000003a
IPv6 outbound checksum enabled: 0x00000000
IPv4 inbound checksum enabled: 0x00000000
IPv6 inbound checksum enabled: 0x00000000
IPv4 Multicast Address:                  MAC Address:
-----------------------                  ------------
224.0.0.1                                01:00:5e:00:00:01
IPv6 Address:                             IPA Flags:
-----------------------                  ------------
fe80::6cae:8b00:748:b68                   0x00000000
IPv6 Multicast Address:                  MAC Address:
-----------------------                  ------------
ff01::1                                 33:33:00:00:00:01
ff02::1                                 33:33:00:00:00:01
ff02::1:ff48:b68                       33:33:ff:48:0b:68
ff02::1:3                               33:33:00:01:00:03
```

This example shows how the binary output from `qethqoat` can be processed in another tool. Here it is displayed in a hexdump viewer:

```
# qethqoat -s 0 enccw0.0.f500
Supported Scope mask: 0x00000001
Supported Descriptor hdr types: 0x0001070f
```

This example shows how the binary output from `qethqoat` can be processed in another tool. Here it is displayed in a hexdump viewer:
# qethqoat -r enccw0.0.f500 | hexdump
00000000 0158 0000 0000 0000 0000 0000 0101 0000 0000
0000010 0000 0001 0000 0000 0000 0000 0000 0000 0000
0000020 0004 0050 0001 0000 0000 0000 0000 d7c8 4040
0000030 0120 0094 001a 643b 8a22 0000 0000 0000
0000040 e102 0002 0000 0004 0001 0000 0000 0800 0000
0000050 0100 0480 0000 0766 0000 0000 0000 0000 0000
0000060 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000070 0008 0060 0001 0000 0000 0000 0000 d5c8 4040
0000080 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000090 0000 0000 0000 0000 0000 0000 0000 0000 0000
00000a0 0021 5c60 0000 001a 0000 0000 0000 001a 0000
00000b0 0000 0000 0000 0000 0000 0000 0000 0000 0000
00000c0 0002 0000 0000 0000 0000 0000 0000 0000 0000
00000d0 0010 0030 0001 0000 0000 0000 0000 c4c8 f6d4
00000e0 0000 0002 0000 0000 0000 0001 0000 0100 0000
00000f0 0001 0001 0000 0000 0000 0000 0000 0000 0000
0000100 e000 0001 0100 5e00 0001 0000 0000 0000 0000
0000110 0010 0030 0001 0000 0000 0000 c4c8 f6d4
0000120 0000 0008 0000 0000 0000 0001 0000 0018 0000
0000130 0001 0001 0000 0000 0000 0000 0000 0000 0000
0000140 ff02 0000 0000 0000 0000 0000 0000 0000 0001
0000150 3333 0000 0001 0000
0000158
**scsi_logging_level - Set and get the SCSI logging level**

Use the `scsi_logging_level` command to create, set, or get the SCSI logging level.

The SCSI logging feature is controlled by a 32-bit value – the SCSI logging level. This value is divided into 3-bit fields that describe the log level of a specific log area. Due to the 3-bit subdivision, setting levels or interpreting the meaning of current levels of the SCSI logging feature is not trivial. The `scsi_logging_level` script helps with both tasks.

**scsi_logging_level syntax**

Where:

- `-a <level>` or `--all <level>` specifies value for all SCSI_LOG fields.
- `-E <level>` or `--error <level>` specifies SCSI_LOG_ERROR.
- `-T <level>` or `--timeout <level>` specifies SCSI_LOG_TIMEOUT.
- `-S <level>` or `--scan <level>` specifies SCSI_LOG_SCAN.
- `-M <level>` or `--midlevel <level>` specifies SCSI_LOG_MLQUEUE and SCSI_LOG_MLCOMPLETE.
- `--mlqueue <level>` specifies SCSI_LOG_MLQUEUE.
- `--mlcomplete <level>` specifies SCSI_LOG_MLCOMPLETE.
- `-L <level>` or `--lowlevel <level>` specifies SCSI_LOG_LLQUEUE and SCSI_LOG_LLCOMPLETE.
- `--llqueue <level>` specifies SCSI_LOG_LLQUEUE.
- `--llcomplete <level>` specifies SCSI_LOG_LLCOMPLETE.
- `-H <level>` or `--hlqueue <level>` specifies SCSI_LOG_HLQUEUE.
- `--hlcomplete <level>` specifies SCSI_LOG_HLCOMPLETE.
- `-I <level>`
--llcomplete <level>
specifies SCSI_LOG_LLCOMPLETE.
-H or --highlevel <level>
specifies SCSI_LOG_HLQUEUE and SCSI_LOG_HLCOMPLETE.
--hlqueue <level>
specifies SCSI_LOG_HLQUEUE.
--hlcomplete <level>
specifies SCSI_LOG_HLCOMPLETE.
-I <level> or --ioctl <level>
specifies SCSI_LOG_IOCTL.
-v or --version
displays version information.
-h or --help
displays help text.
-s or --set
creates and sets the logging level as specified on the command line.
-g or --get
gets the current logging level.
-c or --create
creates the logging level as specified on the command line.

You can specify several SCSI_LOG fields by using several options. When multiple options specify the same SCSI_LOG field, the most specific option has precedence.

Examples

- This command prints the logging word of the SCSI logging feature and each logging level.

```
#> scsi_logging_level -g
Current scsi logging level:
  dev.scsi.logging_level = 0
  SCSI_LOG_ERROR=0
  SCSI_LOG_TIMEOUT=0
  SCSI_LOG_SCAN=0
  SCSI_LOG_MLQUEUE=0
  SCSI_LOG_MLCOMPLETE=0
  SCSI_LOG_LLQUEUE=0
  SCSI_LOG_LLCOMPLETE=0
  SCSI_LOG_HLQUEUE=0
  SCSI_LOG_HLCOMPLETE=0
  SCSI_LOG_IOCTL=0
```

- This command is useful to find issues with LUN discovery and SCSI error handling (recovery), such caused by dirty fibre optics, and has only negligible impact on regular I/O:
```bash
#> scsi_logging_level -s --mlcomplete 1 -T 7 -E 5 -S 7 -I 0 -a 0
New scsi logging level:
dev.scsi.logging_level = 4605
SCSI_LOG_ERROR=5
SCSI_LOG_TIMEOUT=7
SCSI_LOG_SCAN=7
SCSI_LOG_MLQUEUE=0
SCSI_LOG_MLCOMPLETE=1
SCSI_LOG_LLQUEUE=0
SCSI_LOG_LLCOMPLETE=0
SCSI_LOG_HLQUEUE=0
SCSI_LOG_HLCOMPLETE=0
SCSI_LOG_IOCTL=0

v
This command sets all logging levels to 3:
```
t ape390_crypt - Manage tape encryption

Use the tape390_crypt command to enable and disable tape encryption for a channel attached tape device. You can also specify key encrypting keys (KEK) by using labels or hashes.

For 3592 tape devices, it is possible to write data in an encrypted format. The encryption keys are stored on an encryption key manager (EKM) server, which can run on any machine with TCP/IP and Java support. The EKM communicates with the tape drive over the tape control unit by using TCP/IP. The control unit acts as a proxy and forwards the traffic between the tape drive and the EKM. This type of setup is called out-of-band control-unit based encryption.

The EKM creates a data key that encrypts data. The data key itself is encrypted with KEKs and is stored in so called external encrypted data keys (EEDKs) on the tape medium.

You can store up to two EEDKs on the tape medium. With two EEDKs, one can contain a locally available KEK and the other can contain the public KEK of the location or company to where the tape is to be transferred. Then, the tape medium can be read in both locations.

When the tape device is mounted, the tape drive sends the EEDKs to the EKM. The EKM tries to unwrap one of the two EEDKs and sends back the extracted data key to the tape drive.

Linux can address KEKs by specifying either hashes or labels. Hashes and labels are stored in the EEDKs.

Note: If a tape is encrypted, it cannot be used for IPL.

Before you begin:

To use tape encryption, you need:

- A 3592 crypto-enabled tape device and control unit that is configured as system-managed encryption.
- A crypto-enabled 3590 channel-attached tape device driver. See Chapter 13, “Channel-attached tape device driver,” on page 205.
tape390_crypt syntax

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><code>-q</code> or <code>--query</code></td>
<td>displays information about the tape's encryption status. If encryption is active and the medium is encrypted, additional information about the encryption keys is displayed.</td>
</tr>
<tr>
<td><code>-e</code> or <code>--encryption</code></td>
<td>sets tape encryption on or off.</td>
</tr>
<tr>
<td><code>-k</code> or <code>--key</code></td>
<td>sets tape encryption keys. You can only specify the <code>-k</code> option if the tape medium is loaded and rewound. While processing the <code>-k</code> option, the tape medium is initialized and all previous data contained on the tape medium is lost. You can specify the <code>-k</code> option twice because the tape medium can store two EEDKs. If you specify the <code>-k</code> option once, two identical EEDKs are stored.</td>
</tr>
<tr>
<td><code>&lt;value&gt;</code></td>
<td>specifies the key encrypting key (KEK), which can be up to 64 characters long. The keywords <code>label</code> or <code>hash</code> specify how the KEK in <code>&lt;value&gt;</code> is to be stored on the tape medium. The default store type is <code>label</code>.</td>
</tr>
<tr>
<td><code>-d</code> or <code>--delimiter</code></td>
<td>specifies the character that separates the KEK in <code>&lt;value&gt;</code> from the store type (<code>label</code> or <code>hash</code>). The default delimiter is “:” (colon).</td>
</tr>
<tr>
<td><code>&lt;char&gt;</code></td>
<td>is a character that separates the KEK in <code>&lt;value&gt;</code> from the store type (<code>label</code> or <code>hash</code>).</td>
</tr>
<tr>
<td><code>-f</code> or <code>--force</code></td>
<td>specifies that no prompt message is to be issued before writing the KEK information and initializing the tape medium.</td>
</tr>
<tr>
<td><code>&lt;node&gt;</code></td>
<td>specifies the device node of the tape device.</td>
</tr>
</tbody>
</table>

Notes:

1. The `-k` or `--key` operand can be specified maximally twice.
Examples

The following scenarios illustrate the most common use of tape encryption. In all examples /dev/ntibm0 is used as the tape device.

Querying a tape device before and after encryption is turned on

This example shows a query of tape device /dev/ntibm0. Initially, encryption for this device is off. Encryption is then turned on, and the status is queried again.

```
tape390_crypt -q /dev/ntibm0
ENCRIPTION: OFF
MEDIUM: NOT ENCRYPTED

tape390_crypt -e on /dev/ntibm0

tape390_crypt -q /dev/ntibm0
ENCRIPTION: ON
MEDIUM: NOT ENCRYPTED
```

Then, two keys are set, one in label format and one in hash format. The status is queried and there is now additional output for the keys.

```
tape390_crypt -k my_first_key:label -k my_second_key:hash /dev/ntibm0
--- ATTENTION! ---
All data on tape /dev/ntibm0 will be lost.
Type "yes" to continue: yes
SUCCESS: key information set.

tape390_crypt -q /dev/ntibm0
ENCRIPTION: ON
MEDIUM: ENCRYPTED
KEY1:
  value: my_first_key
  type: label
  ontape: label
KEY2:
  value: my_second_key
  type: label
  ontape: hash
```

Using default keys for encryption

1. Load the cartridge. If the cartridge is already loaded:
   • Switch off encryption:
     tape390_crypt -e off /dev/ntibm0
   • Rewind:
     mt -f /dev/ntibm0 rewind
2. Switch encryption on:
   tape390_crypt -e on /dev/ntibm0
3. Write data.
Using specific keys for encryption
1. Load the cartridge. If the cartridge is already loaded, rewind:
   `mt -f /dev/ntibm0 rewind`
2. Switch encryption on:
   `tape390_crypt -e on /dev/ntibm0`
3. Set new keys:
   `tape390_crypt -k key1 -k key2 /dev/ntibm0`
4. Write data.

Writing unencrypted data
1. Load the cartridge. If the cartridge is already loaded, rewind:
   `mt -f /dev/ntibm0 rewind`
2. If encryption is on, switch off encryption:
   `tape390_crypt -e off /dev/ntibm0`
3. Write data.

Appending new files to an encrypted cartridge
1. Load the cartridge
2. Switch encryption on:
   `tape390_crypt -e on /dev/ntibm0`
3. Position the tape.
4. Write data.

Reading an encrypted tape
1. Load the cartridge
2. Switch encryption on:
   `tape390_crypt -e on /dev/ntibm0`
3. Read data.
**tape390_display - Display messages on tape devices and load tapes**

Use the `tape390_display` command to show messages on the display unit of a physical tape device, optionally in conjunction with loading a tape.

**tape390_display syntax**

```
tape390_display [-l] [-q] [-t standard [message1] [message2]]
```

**Where:**

- **-l or --load**
  - instructs the tape unit to load the next indexed tape from the automatic tape loader (if installed). Ignored if no loader is installed or if the loader is not in “system” mode. The loader “system” mode allows the operating system to handle tape loads.

- **-t or --type**
  - The possible values have the following meanings:
    - **standard**
      - displays the message or messages until the physical tape device processes the next tape movement command.
    - **load**
      - displays the message or messages until a tape is loaded; if a tape is already loaded, the message is ignored.
    - **unload**
      - displays the message or messages while a tape is loaded; if no tape is loaded, the message is ignored.
    - **reload**
      - displays the first message while a tape is loaded and the second message when the tape is removed. If no tape is loaded, the first message is ignored and the second message is displayed immediately. The second message is displayed until the next tape is loaded.
    - **noop**
      - is intended for test purposes only. It accesses the tape device but does not display the message or messages.

- **-b or --blink**
  - causes `<message1>` to be displayed repeatedly for 2 seconds with a half-second pause in between.

- `<message1>`
  - is the first or only message to be displayed. The message can be up to 8 byte.

- `<message2>`
  - is a second message to be displayed alternately with the first, at 2-second intervals. The message can be up to 8 byte.
```plaintext
<node>
  is a device node of the target tape device.

-q or --quiet
  suppresses all error messages.

-h or --help
  displays help text.

-v or --version
  displays information about the version.

Note:
1. Symbols that can be displayed include:
   
   **Alphabetic characters:**
   A through Z (uppercase only) and spaces. Lowercase letters are converted to uppercase.

   **Numeric characters:**
   0 1 2 3 4 5 6 7 8 9

   **Special characters:**
   @ $ # , / ( ) * & + - = % : _ < > ? ;
   The following are included in the 3490 hardware reference but might not display on all devices: | "

2. If only one message is defined, it remains displayed until the tape device driver next starts to move or the message is updated.

3. If the messages contain spaces or shell-sensitive characters, they must be enclosed in quotation marks.

Examples

The following examples assume that you are using standard devices nodes and not device nodes that are created by udev:

* Alternately display “BACKUP” and “COMPLETE” at 2-second intervals until device /dev/ntibm0 processes the next tape movement command:
  ```
tape390_display BACKUP COMPLETE /dev/ntibm0
  ```

* Display the message “REM TAPE” while a tape is in the physical tape device followed by the message “NEW TAPE” until a new tape is loaded:
  ```
tape390_display --type reload "REM TAPE" "NEW TAPE" /dev/ntibm0
  ```

* Attempts to unload the tape and load a new tape automatically, the messages are the same as in the previous example:
  ```
tape390_display -l -t reload "REM TAPE" "NEW TAPE" /dev/ntibm0
  ```
```
tunedasd - Adjust low-level DASD settings

Use the `tunedasd` command to adjust performance relevant settings and other low-level DASD device settings.

In particular, you can perform these tasks:

- Query and set a DASD's cache mode
- Display and reset DASD performance statistics
- Reserve and release DASD
- Break the lock of an online DASD (to learn how to access a boxed DASD that is not yet online, see "Accessing DASD by force" on page 137)

**Before you begin:** For the performance statistics, data gathering must be turned on by writing "on" to `/proc/dasd/statistics`.

**tunedasd syntax**

```
-tunedasd -h
       -g or --get_cache
       -c <mode> or --cache <mode>
       [-p <chpid>]
       --path_reset_all
       -n <cylinders>
       -Q
       -S
       -L
       -O
       -R
       -P
       -I <row>
```

Where:

- `<node>` specifies a device node for the DASD to which the command is to be applied.

- `-g` or `--get_cache` gets the current caching mode of the storage controller. This option applies to ECKD only.

- `-c <mode>` or `--cache <mode>` sets the caching mode on the storage controller to `<mode>`. This option applies to ECKD only.

Today’s ECKD devices support the following behaviors:

- **normal** for normal cache replacement.
- **bypass** to bypass cache.
- **inhibit** to inhibit cache.
- **sequential** for sequential access.
- **prestage** for sequential prestage.
- **record** for record access.
tunedasd

For details, see IBM TotalStorage Enterprise Storage Server® System/390® Command Reference 2105 Models E10, E20, F10, and F20, SC26-7295.

-\n\n\ncylinders> or --no_cyl \ncylinders>

specifies the number of cylinders to be cached. This option applies to ECKD only.

-Q or --query_reserve

queries the reserve status of the device. The status can be:

none  the device is not reserved.

im\n\n\nlicit  the device is not reserved, but there is a contingent or implicit allegiance to this Linux instance.

other  the device is reserved to another operating system instance.

reserved  the device is reserved to this Linux instance.

For details, see the “Storage Control Reference” of the attached storage server.

This option applies to ECKD only.

-S or --reserve

reserves the device. This option applies to ECKD only.

-L or --release

releases the device. This option applies to ECKD only.

-O or --slock

reserves the device unconditionally. This option applies to ECKD only.

Note: This option is to be used with care as it breaks any existing reserve by another operating system.

-R or --reset_prof

resets the profile information of the device.

-P or --profile

displays a usage profile of the device.

-I <row> or --prof_item <row>

prints the usage profile item that is specified by <row>. <row> can be one of:

reqs  number of DASD I/O requests.
sects  number of 512-byte sectors.
sizes  histogram of sizes.
total  histogram of I/O times.
totsect  histogram of I/O times per sector.
start  histogram of I/O time until sshc.
irq  histogram of I/O time between sshc and irq.
irqsect  histogram of I/O time between sshc and irq per sector.
end  histogram of I/O time between irq and end.
queue  number of requests in the DASD internal request queue at enqueueing.

-p <chpid> or --path_reset <chpid>

resets a channel path <chpid> of a selected device. A channel path might be suspended due to high IFCC error rates or a High Performance FICON failure. Use this option to resume considering the channel path for I/O.

--path_reset_all

resets all channel paths of a selected device. The channel paths might be
Examples

- The following sequence of commands first checks the reservation status of a DASD and then reserves it:

```
# tunedasd -o /dev/dasdzzz
none
# tunedasd -S /dev/dasdzzz
Reserving device </dev/dasdzzz>...
Done.
# tunedasd -o /dev/dasdzzz
reserved
```

- This example first queries the current setting for the cache mode of a DASD with device node /dev/dasdzzzz and then sets it to one cylinder “prestage”.

```
# tunedasd -g /dev/dasdzzzz
normal (0 cyl)
# tunedasd -c prestage -n 2 /dev/dasdzzzz
Setting cache mode for device </dev/dasdzzzz>...
Done.
# tunedasd -g /dev/dasdzzzz
prestage (2 cyl)
```

- In this example two device nodes are specified. The output is printed for each node in the order in which the nodes where specified.

```
# tunedasd -g /dev/dasdzzzz /dev/dasdzzzy
prestage (2 cyl)
normal (0 cyl)
```

- The following command prints the usage profile of a DASD.

```
# tunedasd -P /dev/dasdzzz
19617 dasd I/O requests
with 484136 sectors(512B each)

# configured I/O times (microseconds)

<table>
<thead>
<tr>
<th>100M</th>
<th>512</th>
<th>1M</th>
<th>2M</th>
<th>4M</th>
<th>8M</th>
<th>16M</th>
<th>32M</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

```

- The following command prints a row of the usage profile of a DASD. The output is on a single line as indicated by the (cont...) (... cont) in the illustration:
The following command resets a failed channel path with CHPID 45:

```
# tunedasd -p 45 /dev/dasdc
```
vmcp - Send CP commands to the z/VM hypervisor

Use the **vmcp** command to send control program (CP) commands to the z/VM hypervisor and display the response from z/VM.

The **vmcp** command expects the command line as a parameter and returns the response to stdout. Error messages are written to stderr.

You can issue **vmcp** commands using the */dev/vmcp* device node (see Chapter 36, “z/VM CP interface device driver,” on page 413) or from a command prompt in a terminal session. In both cases, you must load the vmcp module.

**Before you begin:** Ensure that vmcp is loaded by issuing: `modprobe vmcp`.

**vmcp syntax**

```
  vmcp [-h] [-v] [-k] [-b <size>] <command>
```

Where:

- **-h** or **--help**
  displays help information.

- **-v** or **--version**
  displays version information.

- **-k** or **--keepcase**
  preserves the case of the characters in the specified command string. By default, the command string is converted to uppercase characters.

- **-b <size>** or **--buffer <size>**
  specifies the buffer size in bytes for the response from z/VM CP. Valid values are from 4096 (or 4k) up to 1048756 (or 1M). By default, **vmcp** allocates an 8192 byte (8k) buffer. You can use k and M to specify kilo- and megabytes.

- **<command>**
  specifies the command that you want to send to CP.

If the command completes successfully, **vmcp** returns 0. Otherwise, **vmcp** returns one of the following values:

1. CP returned a non-zero response code.
2. The specified buffer was not large enough to hold CP’s response. The command was run, but the response was truncated. You can use the **--buffer** option to increase the response buffer.
3. Linux reported an error to **vmcp**. See the error message for details.
4. The options that are passed to **vmcp** were erroneous. See the error messages for details.
Examples

- To get your user ID issue:
  
  ```
  # vmcp query userid
  ```

- To attach the device 1234 to your guest, issue:

  ```
  # vmcp attach 1234 
  ```

- If you add the following line to `/etc/sudoers`:

  ```
  ALL ALL=NOPASSWD:/sbin/vmcp
  ```

  every user on the system can run the `indicate` command by using:

  ```
  # sudo vmcp indicate
  ```

- If you need a larger response buffer, use the `--buffer` option:

  ```
  # vmcp --buffer=128k q 1-fffe
  ```
Use the `vmur` command to work with z/VM spool file queues.

The `vmur` command provides these main functions:

**Receive**
Read data from the z/VM reader file queue. The command performs the following steps:
- Places the reader queue file to be received at the top of the queue.
- Changes the reader queue file attribute to NOHOLD.
- Closes the z/VM reader after the file is received.

The `vmur` command detects z/VM reader queue files in:
- VMDUMP format as created by CP VMDUMP.
- NETDATA format as created by CMS SENDFILE or TSO XMIT.

**Punch or print**
Write data to the z/VM punch or printer file queue and transfer it to another user’s virtual reader, optionally on a remote z/VM node. The data is sliced up into 80-byte or 132-byte chunks (called records) and written to the punch or printer device. If the data length is not an integer multiple of 80 or 132, the last record is padded.

**List**
Display detailed information about one or all files on the specified spool file queue.

**Purge**
Remove one or all files on a spool file queue.

**Order**
Position a file at the top of a spool file queue.

**Before you begin:** To use the receive, punch, and print functions, the `vmur` device driver must be loaded and the corresponding unit record devices must be set online.

**Serialization**

The `vmur` command provides strict serialization of all its functions other than list, which does not affect a file queue’s contents or sequence. Thus concurrent access to spool file queues is blocked to prevent unpredictable results or destructive conflicts.

For example, this serialization prevents a process from issuing `vmur purge -f` while another process is running `vmur receive 1234`. However, `vmur` is not serialized against concurrent CP commands that are issued through `vmcp`: if one process is running `vmur receive 1234` and another process issues `vmcp purge rdr 1234`, then the received file might be incomplete. To avoid such unwanted effects, always use `vmur` to work with z/VM spool file queues.

**Spooling options**

With the `vmur` command, you can temporarily override the z/VM settings for the CLASS, DEST, FORM, and DIST spooling options for virtual unit record devices. The `vmur` command restores the original settings before it returns control.

For details about the spooling options, see the z/VM product information. In particular, see the sections about the z/VM CP SPOOL, QUERY VIRTUAL RDR, QUERY VIRTUAL PUN, and QUERY VIRTUAL PRT commands in z/VM CP
Where these are the main command options:

**re or receive**
receives a file from the z/VM reader queue.

**pun or punch**
writes to the z/VM punch queue.

**pr or print**
writes to the z/VM printer queue.
### vmur

**li or list**
lists information about one or all files on a z/VM spool file queue.

**pur or purge**
purges one or all files from a z/VM spool file queue.

**or or order**
places a file on a z/VM spool file queue at the top of the queue.

**Note:** The short forms that are given for receive, punch, print, list, purge, and order are the shortest possible abbreviations. In keeping with z/VM style, you can abbreviate commands by dropping any number of letters from the end of the full keywords until you reach the short form. For example, `vmur re`, `vmur rec`, or `vmur rece` are all equivalent.

The remaining specifications are listed alphabetically by switch. Variable specifications that do not require a switch are listed first.

**<file>**
specifies a file, in the Linux file system, with data to be punched or printed. If this specification is omitted, the data is read from standard input.

**<outfile>**
specifies a file, in the Linux file system, to receive data from the reader spool file. If neither a file name nor `--stdout` are specified, the name and type of the spool file to be received (see the NAME and TYPE columns in `vmur list` output) are used to build an output file name of the form `<name>.<type>`. If the spool file to be received is an unnamed file, an error message is issued.

Use the `--force` option to overwrite existing files without a confirmation prompt.

**<spoolid>**
specifies the spool ID of a file on the z/VM reader, punch, or printer queue. Spool IDs are decimal numbers in the range 0-9999.

For the list or purge function: omitting the spool ID lists or purges all files in the queue.

**-b <sep>,<pad> or --blocked <sep>,<pad>**
receives or writes a file in blocked mode, where `<sep>` specifies the separator and `<pad>` specifies the padding character in hexadecimal notation. Example:

```bash
vmur rec 7 -b 0x25,0x40 -O
```

Use this option to use character sets other than IBM037 and ISO-8859-1 for conversion.

- For the receive function: All trailing padding characters are removed from the end of each record that is read from the virtual reader and the separator character is inserted afterward. The receive function's output can be piped to `iconv` by using the appropriate character sets. Example:

  ```bash
  # vmur rec 7 -b 0x25,0x40 -O | iconv -f EBCDIC-US -t ISO-8859-1 > myfile
  ```

- For the punch or print function: The separator is used to identify the line end character of the file to punch or print. If a line has fewer characters than the record length of the used unit record device, the residual of the record is filled up with the specified padding byte. If a line exceeds the record size, an error is printed. Example:
-c or --convert
converts a VMDUMP spool file into a format appropriate for further analysis with crash.

-C <class> or --class <class>
specifies a spool class.
- For the receive function: The file is received only if it matches the specified class.
- For the purge function: Only files with the specified class are purged.
- For the punch or printer function: Sets the spool class for the virtual reader or virtual punch device. Output files inherit the spool class of the device.

The class is designated by a single alphanumeric character. For receive, it can also be an asterisk (*) to match all classes. Lowercase alphabetic characters are converted to uppercase.

See also "Spooling options" on page 631.

--dest <device>
sets the destination device for spool files that are created on the virtual punch or printer device. The value can be ANY, OFF, or it must be a valid device as defined on z/VM.

See also "Spooling options" on page 631.

-d or --device
specifies the device node of the virtual unit record device.
- If omitted in the receive function, /dev/vmrdr-0.0.000c is assumed.
- If omitted in the punch function, /dev/vmpun-0.0.000d is assumed.
- If omitted in the print function, /dev/vmprt-0.0.000e is assumed.

--dist <distcode>
sets the distribution code for spool files that are created on the virtual punch or printer device. The value can be an asterisk (*), OFF, or it must be a valid distribution code as defined on z/VM.

OFF and * are equivalent. Both specifications reset the distribution code to the value that is set in the user directory.

See also "Spooling options" on page 631.

-f or --force
suppresses confirmation messages.
- For the receive function: overwrites an existing output file without prompting for a confirmation.
- For the punch or print option: automatically converts the Linux input file name to a valid spool file name without any error message.
- For the purge function: purges the specified spool files without prompting for a confirmation.

--form <form_name>
sets the form name for spool files that are created on the virtual punch or printer device. The value can be OFF, to use the system default, or it must be a valid z/VM form name.

See also "Spooling options" on page 631.
-h or --help
displays help information for the command. To view the man page, enter man
vmur.

-H or --hold
keeps the spool file to be received in the reader queue. If omitted, the spool file
is purged after it is received.

-n <node> or --node <node>
specifies the node name of the z/VM system to which the data is to be
transferred. Remote Spooling Communications Subsystem (RSCS) must be
installed on the z/VM systems and the specified node must be defined in the
RSCS machine’s configuration file.

The default node is the local z/VM system. The node option is valid only with
the -u option.

-N <name>.<type> or --name <name>.<type>
specifies a name and, optionally, a type for the z/VM spool file to be created
by the punch or print option. To specify a type after the file name, enter a
period followed by the type. For example:

```
# vmur pun -r /boot/parmfile -N myname.mytype
```

Both the name and the type must comply with z/VM file name rules, for
example, they must be 1 - 8 characters long.

If omitted, a spool file name is generated from the Linux input file name, if
applicable.

Use the --force option to suppress warning messages about automatically
generated file names or about specified file names that do not adhere to the
z/VM file naming rules.

-o or --stdout
writes the reader file content to standard output.

-q or --queue
specifies the z/VM spool file queue to be listed, purged, or ordered. If omitted,
the reader file queue is assumed.

-r or --rdr
transfers a punch or print file to a reader.

-t or --text
converts the encoding between EBCDIC and ASCII according to character sets
IBM037 and ISO-8859-1.

- For the receive function: receives the reader file as text file. That is, it
  converts EBCDIC to ASCII and inserts an ASCII line feed character (0x0a)
  for each input record that is read from the z/VM reader. Trailing EBCDIC
  blanks (0x40) in the input records are stripped.
- For the punch or print function: punches or prints the input file as text file.
  That is, converts ASCII to EBCDIC and pads each input line with trailing
  blanks to fill up the record. The record length is 80 for a punch and 132 for a
  printer. If an input line length exceeds 80 for punch or 132 for print, an error
  message is issued.

The --text and the --blocked attributes are mutually exclusive.
-u <user> or --user <user>
specifies the z/VM user ID to whose reader the data is to be transferred. If
omitted, the data is transferred to your own machine’s reader. The user option
is valid only with the -r option.

-v or --version
displays version information.

Examples
These examples illustrate common scenarios for unit record devices.

In all examples the following device nodes are used:
* /dev/vmrdr-0.0.000c as virtual reader.
* /dev/vmpun-0.0.000d as virtual punch.

The vmur commands access the reader device, which has to be online. To set it
online, it needs to be freed from cio_ignore. Example:

```
# cio_ignore -r c
# chccwdev -e c
Setting device 0.0.000c online
Done
```

Besides the vmur device driver and the vmur command, these scenarios require
that the vmcp and vmconvert commands from the s390utils package are available.

Creating and reading a guest memory dump
You can use the vmur command to read a guest memory dump that was created,
for example, with the vmcp command.

Procedure
1. Produce a memory dump of the z/VM guest virtual machine memory:

```
# vmcp vmdump
```

Depending on the memory size this command might take some time
to complete.

2. List the spool files for the reader to find the spool ID of the dump file,
VMDUMP. In the example, the spool ID of VMDUMP is 463.

```
# vmur 1;
```

```
ORIGINID  FILE   CLASS  RECORDS  CPY  HOLD DATE    TIME   NAME    TYPE   DIST
T6360025  0463   VDMP   00020222  001  NONE 06/11 15:07:42 VMDUMP FILE T6360025
```

3. Read and convert the VMDUMP spool file to a file in the current working
directory of the Linux file system:

```
# vmur rec 463 -c linux_dump
```

Using FTP to receive and convert a dump file:

Use the --convert option together with the --stdout option to receive a VMDUMP
spool file straight from the z/VM reader queue, convert it, and send it to another
host with FTP.
Procedure
1. Establish an FTP session with the target host and log in.
2. Enter the FTP command binary.
3. Enter the FTP command:
   
   ```
   put "vmur re <spoolid> -c -O" <filename_on_target_host>
   ```

Log and read the z/VM guest virtual machine console
You can use the `vmur` command to read a console transcript that has been spooled; for example, with the `vmcp` command.

Procedure
1. Begin console spooling:
   
   ```
   # vmcp sp cons start
   ```
2. Produce output to the z/VM console (for example, with CP TRACE).
3. Stop console spooling, close the file with the console output, and transfer the file to the reader queue. In the resulting CP message, the spool ID follows the FILE keyword. In the example, the spool ID is 398:
   
   ```
   # vmcp sp cons stop close \* rdr
   RDR FILE 0398 SENT FROM T6360025 CON WAS 0398 RECS 1872 CPY 001 T NOHOLD NOKEEP
   ```
4. Read the file with the console output into a file in the current working directory on the Linux file system:
   
   ```
   # vmur re -t 398 linux_cons
   ```

Preparing the z/VM reader as an IPL device for Linux
You can use the `vmur` command to transfer all files for booting Linux to the z/VM reader. You can also arrange the files such that the reader can be used as an IPL device.

Procedure
1. Send the kernel parameter file, `parmfile`, to the z/VM punch device and transfer the file to the reader queue. The resulting message shows the spool ID of the parameter file.
   
   ```
   # vmur pun -r /boot/parmfile
   Reader file with spoolid 0465 created.
   ```
2. Send the kernel image file to the z/VM punch device and transfer the file to the reader queue. The resulting message shows the spool ID of the kernel image file.
   
   ```
   # vmur pun -r /boot/vmlinuz -N image
   Reader file with spoolid 0466 created.
   ```
3. Optional: Check the spool IDs of `image` and `parmfile` in the reader queue. In this example, the spool ID of `parmfile` is 465 and the spool ID of `image` is 466.
4. Move image to the first and parmfile to the second position in the reader queue:

   # vmur or 465
   # vmur or 466

5. Configure the z/VM reader as the re-IPL device:

   # echo 0.0.000c > /sys/firmware/reipl/ccw/device

6. Boot Linux from the z/VM reader:

   # reboot

Sending a file to different z/VM guest virtual machines
You can use the vmur command to send files to other z/VM guest virtual machines.

About this task
This scenario describes how to send a file called lnxprofile.exec from the file system of an instance of Linux on z/VM to other z/VM guest virtual machines. For example, lnxprofile.exec could contain the content of a PROFILE EXEC file with CP and CMS commands to customize z/VM guest virtual machines for running Linux.

Procedure
1. Send lnxprofile.exec to two z/VM guest virtual machines: z/VM user ID t2930020 at node boet2930 and z/VM user ID t6360025 at node boet6360.

   vmur pun lnxprofile.exec -t -r -u t2930020 -n boet2930 -N PROFILE
   vmur pun lnxprofile.exec -t -r -u t6360025 -n boet6360 -N PROFILE

2. Log on to t2930020 at boet2930, IPL CMS, and issue the CP command:

   QUERY RDR ALL

   The command output shows the spool ID of PROFILE in the FILE column.

3. Issue the CMS command:

   RECEIVE <spoolid> PROFILE EXEC A (REPL)

   In the command, <spoolid> is the spool ID of PROFILE found in step 2.

4. Repeat steps 2 and 3 for t6360025 at boet6360.

Sending a file to a z/VSE instance
You can use the vmur command to send files to a z/VSE instance.
Procedure

To send lserv.job to user ID vseuser at node vse01sys, issue:

```
vmur pun lserv.job -t -r vseuser -n vse01sys -N LSERV
```
Use the `zdsfs` command to mount z/OS DASDs as a Linux file system.

The `zdsfs` file system translates the z/OS data sets, which are stored on the DASDs in records of arbitrary or even variable size, into Linux semantics.

Through the `zdsfs` file system, applications on Linux can read z/OS physical sequential data sets (PS) and partitioned data sets (PDS) on the DASD. In the Linux file system, physical sequential data sets are represented as files. Partitioned data sets are represented as directories that contain the PDS members as files. Other z/OS data set formats, such as extended format data sets or VSAM data sets, are not supported. `zdsfs` is optimized for sequential read access.

The `zdsfs` file system requires the FUSE library. Red Hat Enterprise Linux automatically installs this library. The `zdsfs` command is available from the `s390utils-zdsfs` RPM.

**Attention:**

- To avoid data inconsistencies, set the DASDs offline in z/OS before you mount them in Linux.
- Through the `zdsfs` file system, the whole DASDs are accessible to Linux, but the access is not controlled by z/OS auditing mechanisms.
  
  To avoid security problems, you might want to dedicate the z/OS DASDs only for providing data for Linux.

Per default, only the Linux user who mounts the `zdsfs` file system has access to it.

**Tip:** If you want to grant a user group access to the `zdsfs` file system, mount it with the fuse options `default_permissions`, `allow_other`, and `gid`.

To unmount file systems that you mounted with `zdsfs`, you can use `fusermount`, whether root or non-root user. See the `fusermount` man page for details.

See `z/OS DFSMS Using Data Sets, SC26-7410` for more information about z/OS data sets.

### Before you begin:

- The raw-track access mode of the DASD must be enabled.
  
  Make sure that the DASD is set offline when you enable the raw-track access mode.
  
  See "Accessing full ECKD tracks" on page 149 for details.

- The DASD must be online.

**Tip:** You can use the `chccwdev` command to enable the raw-track access mode and set the device online afterward in one step.

Set the DASD offline in z/OS before you set it online in Linux.

- You must have the appropriate read permissions for the device node.
zdsfs syntax

```
  zdsfs [ <zdsfs-options> ] [ <fuse-options> ] -l <file-name> <node-list>
  [ <mount-point> ]
```

where:

<zdsfs-options>

- **-o ignore_incomplete**
  
  Represents all complete data sets in the file system, even if there are incomplete data sets. Incomplete data sets are not represented.

  In z/OS, data sets might be distributed over different DASDs. For each incomplete data set, a warning message is issued to the standard error stream. If there are incomplete data sets and this option is not specified, the zdsfs command returns with an error.

- **-o rdw**
  
  Keeps record descriptor words (RDWs) of data sets that are stored by using the z/OS concept of variable record lengths.

- **-o tracks=**
  
  Specifies the track buffer size in tracks. The default is 128 tracks.

  zdsfs allocates a track buffer of \(<n>\times 120\) KB for each open file to store and extract the user data. Increasing the track buffer size might improve your system performance.

- **-o seekbuffer=**
  
  Sets the maximum seek history buffer size in bytes. The default is 1,048,576 B.

  zdsfs saves offset information about a data set in the seek history buffer to speed up the performance of a seek operation.

- **-o check_host_count**
  
  Checks the host-access open count to ensure that the device is not online to another operating system instance. The operation is canceled if another operating system instance is accessing the volume.

<fuse-options>

Options for FUSE. The following options are supported by the zdsfs command. To use an option, it must also be supported by the version of FUSE that is installed.

- **-d or -o debug**
  
  Enables debug output (implies -f).

- **-f**
  
  Runs the command as a foreground operation.

- **-o allow_other**
  
  Allows access to other users.

- **-o allow_root**
  
  Allows access to root.
-o nonempty
  allows mounts over files and non-empty directories.

- o default_permissions
  enables permission checking by the kernel.

- o max_read=<n>
  sets maximum size of read requests.

- o kernel_cache
  caches files in the kernel.

- o [no]auto_cache
  enables or disables caching based on modification times.

- o umask=<mask>
  sets file permissions (octal).

- o uid=<n>
  sets the file owner.

- o gid=<n>
  sets the file group.

- o max_write=<n>
  sets the maximum size of write requests.

- o max_readahead=<n>
  sets the maximum readahead value.

- o async_read
  performs reads asynchronously (default).

- o sync_read
  performs reads synchronously.

<node-list>
  one or more device nodes for the DASDs, separated by blanks.

<file-name>
  a file that contains a node list.

<mount-point>
  the mount point in the Linux file system where you want to mount the z/OS data sets.

-h or --help
  displays help information for the command. To view the man page, enter man zdsfs.

-v or --version
  displays version information for the command.

File characteristics

There are two ways to handle the z/OS characteristics of a file:
- The file metadata.txt:

  The metadata.txt file is in the root directory of the mount point. It contains one row for each file or directory, where:

  dsn
    specifies
    - the name of the file in the form <file-name> for z/OS physical sequential data sets.
the name of the directory in the form `<directory-name>`, and the name of a file in that directory in the form `<directory-name>(<file-name>)` for z/OS partitioned data sets.

**dsorg**

specifies the organization of the file. The organization is PO for a directory, and PS for a file.

**lrecl**

specifies the record length of the file.

**recfm**

specifies the z/OS record format of the file. Supported record formats are: V, F, U, B, S, A, and M.

**Example:**

```
dsn=FOOBAR.TESTF.TXT,recfm=FB,lrecl=80,dsorg=PS
dsn=FOOBAR.TESTVB.TXT,recfm=V8,lrecl=100,dsorg=PS
dsn=FOOBAR.PDSF.DAT,recfm=F,lrecl=80,dsorg=PO
dsn=FOOBAR.PDSF.DAT(TEST1),recfm=F,lrecl=80,dsorg=PS
dsn=FOOBAR.PDSF.DAT(TEST2),recfm=F,lrecl=80,dsorg=PS
dsn=FOOBAR.PDSF.DAT(TEXT3),recfm=F,lrecl=80,dsorg=PS
```

**Extended attributes:**

**user.dsorg**

specifies the organization of the file.

**user.lrecl**

specifies the record length of the file.

**user.recfm**

specifies the z/OS record format of the file.

You can use the following system calls to work with extended attributes:

**listxattr**

to list the current values of all extended attributes.

**getxattr**

to read the current value of a particular extended attribute.

You can use these system calls through the `getfattr` command. For more information, see the man pages of these commands and of the listxattr and getxattr system calls.

**Examples**

- Enable the raw-track access mode of DASD device 0.0.7000 and set the device online afterward:

  ```
  # chccwdev -a raw_track_access=1 -e 0.0.7000
  ```

- Mount the partitioned data set on the DASDs represented by the file nodes `/dev/dasde` and `/dev/dasdf` at `/mnt`:

  ```
  # zdfs /dev/dasde /dev/dasdf /mnt
  ```

- As user “myuser”, mount the partitioned data set on the DASD represented by the file node `/dev/dasde` at `/home/myuser/mntzos`:
  
  - Access the mounted file system exclusively:

    ```
    # zdfs /dev/dasde /home/myuser/mntzos
    ```
– Allow the root user to access the mounted file system:

```
# zdsfs -o allow_root /dev/dasde /home/myuser/mntzos
```

The `ls` command does not reflect these permissions. In both cases, it shows:

```
# ls -al /home/myuser/mntzos
total 121284
dr-xr-x--- 2 root  root  0 Dec  3 15:54 .
drwx------ 3 myuser myuser 4096 Dec  3 15:51 ...
-r-r-r---- 1 root  root  2833200 Jun 27 2012 EXPORT.BIN1.DAT
-r-r-r---- 1 root  root  2833200 Jun 27 2012 EXPORT.BIN2.DAT
-r-r-r---- 1 root  root  2833200 Jun 27 2012 EXPORT.BIN3.DAT
-r-r-r---- 1 root  root  2833200 Feb 14 2013 EXPORT.BIN4.DAT
dr-xr-x--- 2 root  root  13599360 Aug 9 2012 EXPORT.PDS1.DAT
dr-xr-x--- 2 root  root  13599360 Aug 9 2012 EXPORT.PDS2.DAT
dr-xr-x--- 2 root  root  13599360 Aug 9 2012 EXPORT.PDS3.DAT
dr-xr-x--- 2 root  root  13599360 Aug 9 2012 EXPORT.PDS4.DAT
-r-r-r---- 1 root  root  55247400 Aug 9 2012 EXPORT.PDS4.DAT
```

As root user, mount the partitioned data set on the DASD represented by the file node `/dev/dasde` at `/mnt` on behalf of the user ID “myuser” (UID=1002), and permit the members of the group ID “zosimport” (GID=1002) file access:

```
# zdsfs /dev/dasde /mnt -o uid=1002,gid=1002,allow_other,default_permissions
```

The `ls` command indicates the owner “myuser” and the access right for group “zosimport”:

```
$ ls -al /mnt
total 121284
dr-xr-x--- 2 myuser zosimport 0 Dec  3 14:22 .
drwxr-xr-x 23 root  root 4096 Dec  3 13:59 ..
-r-r-r---- 1 myuser zosimport 981 Dec  3 14:22 metadata.txt
-r-r-r---- 1 myuser zosimport 2833200 Jun 27 2012 EXPORT.BIN1.DAT
-r-r-r---- 1 myuser zosimport 2833200 Jun 27 2012 EXPORT.BIN2.DAT
-r-r-r---- 1 myuser zosimport 2833200 Jun 27 2012 EXPORT.BIN3.DAT
-r-r-r---- 1 myuser zosimport 2833200 Feb 14 2013 EXPORT.BIN4.DAT
dr-xr-x--- 2 myuser zosimport 13599360 Aug 9 2012 EXPORT.PDS1.DAT
dr-xr-x--- 2 myuser zosimport 13599360 Aug 9 2012 EXPORT.PDS2.DAT
dr-xr-x--- 2 myuser zosimport 13599360 Aug 9 2012 EXPORT.PDS3.DAT
dr-xr-x--- 2 myuser zosimport 13599360 Aug 9 2012 EXPORT.PDS4.DAT
```

Unmount the partitioned data set that is mounted at `/mnt`:

```
# fusermount -u /mnt
```

Show the extended attributes of a file, `FB.XMP.TXT`, on a z/OS DASD that is mounted on `/mnt`:

```
# getfattr -d /mnt/FB.XMP.TXT
```

Show the extended attributes of all files on a z/OS DASD that is mounted on `/mnt`:

```
# cat /mnt/metadata.txt
```

---

**zdsfs**

- Allow the root user to access the mounted file system:

```
# zdsfs -o allow_root /dev/dasde /home/myuser/mntzos
```

The `ls` command does not reflect these permissions. In both cases, it shows:

```
# ls -al /home/myuser/mntzos
```

```
```

- As root user, mount the partitioned data set on the DASD represented by the file node `/dev/dasde` at `/mnt` on behalf of the user ID “myuser” (UID=1002), and permit the members of the group ID “zosimport” (GID=1002) file access:

```
# zdsfs /dev/dasde /mnt -o uid=1002,gid=1002,allow_other,default_permissions
```

The `ls` command indicates the owner “myuser” and the access right for group “zosimport”:

```
$ ls -al /mnt
```

```
```

- Unmount the partitioned data set that is mounted at `/mnt`:

```
# fusermount -u /mnt
```

- Show the extended attributes of a file, `FB.XMP.TXT`, on a z/OS DASD that is mounted on `/mnt`:

```
# getfattr -d /mnt/FB.XMP.TXT
```

- Show the extended attributes of all files on a z/OS DASD that is mounted on `/mnt`:

```
# cat /mnt/metadata.txt
```
zkey - Generate, validate, and re-encipher secure AES keys

Use the zkey command to generate, validate, and re-encipher secure AES keys for Crypto Express CCA co-processors.

The zkey command is used in the context of a cryptographic domain of a Crypto Express adapter in CCA co-processor mode. At least one domain must have an AES master key configured. Such a domain of a cryptographic adapter maintains three registers for master keys: The OLD, the NEW, and the CURRENT. In the following description, OLD, NEW, and CURRENT refer to the master key registers of the applicable domain.

Note: If multiple domains on potentially multiple CCA co-processors are available, the zkey command might use any of those domains, and hence, adapters.

Prerequisites

- The zkey command requires access to a cryptographic domain that has a configured AES master key.
- The zkey command requires the pkey kernel module, see Chapter 42, “Protected key device driver,” on page 447.
- The reencipher command requires the IBM CCA Host Library (libcsulcca.so). The IBM CCA host library is part to the CCA software package, which you can obtain from www.ibm.com/security/cryptocards/pciecc2/lonzsoftware.shtml

Syntax

zkey syntax

Where:

gen <secure_key_file> or generate <secure_key_file>
generates a new secure AES key either randomly within the CCA cryptographic adapter, or from a clear AES key specified as input.

-k or --keybits
specifies the size of the AES key to be generated in bits. Valid sizes are 128, 192, and 256 bits. Secure keys for use with the XTS cipher mode can only use keys of 128 or 256 bits. The default is 256 bits.

-x or --xts
generates a secure AES key for the XTS cipher mode. A secure AES key for the XTS cipher mode consist of two concatenated secure keys.

-c <clear_key_file> or --clearkey <clear_key_file>
specifies a file path that contains the clear AES key in binary form. If the --keybits option is omitted, the size of the specified file determines the size of
the AES key. If the --keybits option is specified, the size of the specified file must match the specified key size. Valid file sizes are 16, 24, or 32 bytes, and for the XTS cipher mode 32 or 64 bytes.

val <secure_key_file> or validate <secure_key_file>
checks whether the specified file contains a valid secure key. It also displays the attributes of the secure key, such as key size, whether it is a secure key that can be used for the XTS cipher mode, and the master key register with which the secure key is enciphered.

re <secure_key_file> or reencipher <secure_key_file>
re-enciphers an existing secure key with a new master key. A secure key must be re-enciphered when the master key of the CCA cryptographic adapter changes.

-n or --to-new
re-enciphers a secure key that is currently enciphered with the master key in the CURRENT register with the master key in the NEW register.

-o or --from-old
re-enciphers a secure key that is currently enciphered with the master key in the OLD register with the master key in the CURRENT register.

If both options are specified, a secure key that is currently enciphered with the master key in the OLD register is re-enciphered with the master key in the NEW register.

If both options are omitted, zkey automatically detects whether the secure key is currently enciphered with the master key in the OLD register or with the master key in the CURRENT register.

If currently enciphered with the master key in the OLD register, it is re-enciphered with the master key in the CURRENT register. If it is currently enciphered with the master key in the CURRENT register, it is re-enciphered with the master key in the NEW register.

-f <output_file> or --output <output_file>
specifies the name of the output file to which the re-enciphered secure key is written. If this option is omitted, the re-enciphered secure key is replaced in the file that currently contains the secure key.

-V or --verbose
displays additional information during processing.

-h or --help
displays help information for the command.

-v or --version
displays the version number of zkey.

Examples
- To generate a 256-bit secure AES key and store it in seckey.bin:
  zkey generate seckey.bin

- To generate a 128-bit secure AES key for the XTS cipher mode and store it in seckey.bin:
  zkey generate seckey.bin --keybits 128 --xts

- To generate a secure AES key from the clear key in file clearkey.bin and store it in seckey.bin:
zkey

```bash
zkey generate seckey.bin --clearkey clearkey.bin
```

- To re-encipher the secure key in `seckey.bin`, which is currently enciphered with the master key in the OLD register with the master key in the CURRENT register, and replace the secure key in `seckey.bin` with the re-enciphered key:
  ```bash
  zkey generate seckey.bin --from-old
  ```

- To re-encipher the secure key in `seckey.bin`, which is currently enciphered with the master key in the CURRENT register, with the master key in the NEW register, and save the re-enciphered secure key to `seckey2.bin`:
  ```bash
  zkey reencipher seckey.bin --to-new --output seckey2.bin
  ```

- To validate the secure key in `seckey.bin` and display its attributes:
  ```bash
  zkey validate seckey.bin
  ```

For more details about command usage, see the man page.
znetconf

znetconf - List and configure network devices

Use the znetconf command to list, configure, add, and remove network devices.

The znetconf command:

- Lists potential network devices.
- Lists configured network devices.
- Automatically configures and adds network devices.
- Removes network devices.

For automatic configuration, znetconf first builds a channel command word (CCW) group device from sensed CCW devices. It then configures any specified option through the sensed network device driver and sets the new network device online.

During automatic removal, znetconf sets the device offline and removes it.

Attention: Removing all network devices might lead to complete loss of network connectivity. Unless you can access your Linux instance from a terminal server on z/VM (see How to Set up a Terminal Server Environment on z/VM, SC34-2596), you might require the HMC or a 3270 terminal session to restore the connectivity.

Before you begin: The qeth, ctcm, or lcs device drivers must be loaded. If needed, the znetconf command attempts to load the particular device driver.

znetconf syntax

```
$ znetconf
```

Where:

- **a** or **--add**
  
  configures the network device with the specified device bus-ID. If you specify only one bus ID, the command automatically identifies the remaining bus IDs of the group device. You can enter a list of device bus-IDs that are separated by commas. The znetconf command does not check the validity of the combination of device bus-IDs.

- **d** <driver>

  specifies the device bus-ID of the CCW devices that constitute the network device. If a device bus-ID begins with "0.0.", you can abbreviate it to the final hexadecimal digits. For example, you can abbreviate 0.0.f503 to f503.
-A or --add-all
configures all potential network devices. After you run `znetconf -A`, enter `znetconf -c` to see which devices were configured. You can also enter `znetconf -u` to display devices that were not configured.

-e or --except
omits the specified devices when configuring all potential network devices or removing all configured network devices.

-o <attribute>=<value> or --option <attribute>=<value>
configures devices with the specified sysfs option.

-d <driver name> or --driver <driver name>
configures devices with the specified device driver. Valid values are qeth, lcs, ctc, or ctcm.

-n or --non-interactive
answers all confirmation questions with "Yes".

-r or --remove
removes the network device with the specified device bus-ID. You can enter a list of device bus-IDs that are separated by a comma. You can remove only configured devices as listed by `znetconf -c`.

-R or --remove-all
removes all configured network devices. After successfully running this command, all devices that are listed by `znetconf -c` become potential devices that are listed by `znetconf -u`.

-u or --unconfigured
lists all network devices that are not yet configured.

-c or --configured
lists all configured network devices.

-h or --help
displays short information about command usage. To view the man page, enter `man znetconf`.

-v or --version
displays version information.

If the command completes successfully, `znetconf` returns 0. Otherwise, 1 is returned.

Examples
- To list all potential network devices:

```
# znetconf -u
Device IDs   Type    Card Type    CHPID   Drv.
------------------------------------------
0.0.f500,0.0.f501,0.0.f502 1731/01 OSA (QDIO) 00  qeth
0.0.f503,0.0.f504,0.0.f505 1731/01 OSA (QDIO) 01  qeth
```

- To configure device 0.0.f503:

```
znetconf -a 0.0.f503
```

or

```
znetconf -a f503
```
To configure the potential network device 0.0.f500 with the `layer2` option with the value 0:

```
znetconf -a f500 -o layer2=0
```

To list configured network devices:

```
znetconf -c
```

<table>
<thead>
<tr>
<th>Device IDs</th>
<th>Type</th>
<th>Card Type</th>
<th>Chpid</th>
<th>Drv. Name</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.f500, 0.0.f501, 0.0.f502</td>
<td>Virt.NIC</td>
<td>QDIO</td>
<td>00</td>
<td>qeth</td>
<td>enccw0.0.f500 online</td>
</tr>
<tr>
<td>0.0.f503, 0.0.f504, 0.0.f505</td>
<td>Virt.NIC</td>
<td>QDIO</td>
<td>01</td>
<td>qeth</td>
<td>enccw0.0.f503 online</td>
</tr>
</tbody>
</table>

To remove network device 0.0.f503:

```
znetconf -r 0.0.f503
```

or

```
znetconf -r f503
```

To remove all configured network devices except the devices with bus IDs 0.0.f500 and 0.0.f5f0:

```
znetconf -R -e 0.0.f500 -e 0.0.f5f0
```

To configure all potential network devices except the device with bus ID 0.0.f503:

```
znetconf -A -e 0.0.f503
```
Chapter 54. Selected kernel parameters

You can use kernel parameters that are beyond the scope of an individual device driver or feature to configure Linux in general.

Device driver-specific kernel parameters are described in the setting up section of the respective device driver.

cio_ignore

**cio_ignore - List devices to be ignored**

Use the `cio_ignore=` kernel parameter to list specifications for I/O devices that are to be ignored.

When an instance of Linux on Z boots, it senses and analyzes all available I/O devices. The following applies to ignored devices:

- Ignored devices are not sensed and analyzed. The device cannot be used until it is analyzed.
- Ignored devices are not represented in sysfs.
- Ignored devices do not occupy storage in the kernel.
- The subchannel to which an ignored device is attached is treated as if no device were attached.
- For Linux on z/VM, cio_ignore might hide essential devices such as the console. The console is typically device number 0.0.0009.

See also “Changing the exclusion list” on page 653.

**Format**

```
cio_ignore syntax

[cio_ignore=    all<device_spec>    ,    <device_spec>]

<device_spec>:

[<device_bus_id>]
[<from_device_bus_id>-<to_device_bus_id>]
[ipldev]
[condev]

Where:

**all**
states that all devices are to be ignored.

**<device_bus_id>**
specifies a device. Device bus-IDs are of the form 0.<n>.<devno>, where <n> is a subchannel set ID and <devno> is a device number.

**<from_device_bus_id>-<to_device_bus_id>**
are two device bus-IDs that specify the first and the last device in a range of devices.

**ipldev**
specifies the IPL device. Use this keyword with the ! operator to avoid ignoring the IPL device.
```
condev
specifies the CCW console. Use this keyword with the ! operator to avoid ignoring the console device.

! makes the following term an exclusion statement. This operator is used to exclude individual devices or ranges of devices from a preceding more general specification of devices.

Examples
- This example specifies that all devices in the range 0.0.b100 through 0.0.b1ff, and the device 0.0.a100 are to be ignored.
  cio_ignore=0.0.b100-0.0.b1ff,0.0.a100
- This example specifies that all devices are to be ignored.
  cio_ignore=all
- This example specifies that all devices except the console are to be ignored.
  cio_ignore=all,!condev
- This example specifies that all devices but the range 0.0.b100 through 0.0.b1ff, and the device 0.0.a100 are to be ignored.
  cio_ignore=all,10.0.b100-0.0.b1ff,10.0.a100
- This example specifies that all devices in the range 0.0.1000 through 0.0.1500 are to be ignored, except for devices in the range 0.0.1100 through 0.0.1120.
  cio_ignore=0.0.1000-0.0.1500,10.0.1100-0.0.1120
  This is equivalent to the following specification:
  cio_ignore=0.0.1000-0.0.10ff,0.0.1121-0.0.1500
- This example specifies that all devices in range 0.0.1000 through 0.0.1100 and all devices in range 0.1.7000 through 0.1.7010, plus device 0.0.1234 and device 0.1.4321 are to be ignored.
  cio_ignore=0.0.1000-0.0.1100,0.1.7000-0.1.7010,0.0.1234,0.1.4321

Changing the exclusion list
Use the cio_ignore command or the procfs interface to view or change the list of I/O device specifications that are ignored.

When an instance of Linux on Z boots, it senses and analyzes all available I/O devices. You can use the cio_ignore kernel parameter to list specifications for devices that are to be ignored.

On a running Linux instance, you can view and change the exclusion list through a procfs interface.

After booting Linux you can display the exclusion list by issuing:

```
# cat /proc/cio_ignore
```

To add device specifications to the exclusion list issue a command of this form:

```
# echo add <device_list> > /proc/cio_ignore
```

When you add specifications for a device that is already sensed and analyzed, there is no immediate effect of adding it to the exclusion list. For example, the
cio_ignore

device still appears in the output of the `lscss` command and can be set online. However, if the device later becomes unavailable, it is ignored when it reappears. For example, if the device is detached in z/VM it is ignored when it is attached again.

To make all devices that are in the exclusion list and that are currently offline unavailable to Linux issue a command of this form:

```
# echo purge > /proc/cio_ignore
```

This command does not make devices unavailable if they are online.

To remove device specifications from the exclusion list issue a command of this form:

```
# echo free <device_list> > /proc/cio_ignore
```

When you remove device specifications from the exclusion list, the corresponding devices are sensed and analyzed if they exist. Where possible, the respective device driver is informed, and the devices become available to Linux.

In these commands, `<device_list>` follows this syntax:

```
<device_list>:
  all <device_spec> ...
  <device_spec> ...

<device_spec>:
  <device_bus_id>
  <from_device_bus_id>-<to_device_bus_id>
```

Where the keywords and variables have the same meaning as in “Format” on page 652.

**Ensure device availability**

After the echo command completes successfully, some time might elapse until the freed device becomes available to Linux. To confirm that a device has become available to Linux verify that the sysfs attribute `/sys/bus/ccw/devices/<device_bus_ID>/online` is present.

**Results**

The dynamically changed exclusion list is only taken into account when a device in this list is newly made available to the system, for example after it is defined to
the system. It does not have any effect on setting devices online or offline within Linux.

Examples

- This command removes all devices from the exclusion list.
  
  ```
  # echo free all > /proc/cio_ignore
  ```

- This command adds all devices in the range 0.0.b100 through 0.0.b1ff and device 0.0.a100 to the exclusion list.
  
  ```
  # echo add 0.0.b100-0.0.b1ff,0.0.a100 > /proc/cio_ignore
  ```

- This command lists the ranges of devices that are ignored by common I/O.
  
  ```
  # cat /proc/cio_ignore
  0.0.0000-0.0.a0ff
  0.0.a101-0.0.b0ff
  0.0.b200-0.0.ffff
  ```

- This command removes all devices in the range 0.0.b100 through 0.0.b1ff and device 0.0.a100 from the exclusion list.
  
  ```
  # echo free 0.0.b100-0.0.b1ff,0.0.a100 > /proc/cio_ignore
  ```

- This command removes the device with bus ID 0.0.c104 from the exclusion list.
  
  ```
  # echo free 0.0.c104 > /proc/cio_ignore
  ```

- This command adds the device with bus ID 0.0.c104 to the exclusion list.
  
  ```
  # echo add 0.0.c104 > /proc/cio_ignore
  ```

- This command makes all devices that are in the exclusion list and that are currently offline unavailable to Linux.
  
  ```
  # echo purge > /proc/cio_ignore
  ```
cmma - Reduce hypervisor paging I/O overhead

Use the cmma= kernel parameter to reduce hypervisor paging I/O overhead.

You can use Collaborative Memory Management Assist (CMMA, or "cmm2") on all IBM mainframe systems that Red Hat Enterprise Linux 7.5 supports. With this support, the z/VM control program and guest virtual machines can communicate attributes for specific 4K-byte blocks of guest memory. This exchange of information helps both the z/VM host and the guest virtual machines to optimize their use and management of memory.

Format

```
cmma syntax
```

```
cmma= yes
   on

   cmma= no
   off
```

Examples

This specification disables the CMMA support:
```
cmma=off
```

Alternatively, you can use the following specification to disable the CMMA support:
```
cmma=no
```
fips - Run Linux in FIPS mode

In Federal Information Processing Standard (FIPS) mode, the kernel enforces FIPS 140-2 security standards. For example, in FIPS mode only FIPS 140-2 approved encryption algorithms can be used (see “FIPS restrictions of the hardware capabilities” on page 449).

For more information about FIPS, go to csrc.nist.gov/publications/PubsFIPS.html.

Format

```
fips syntax

fips=0
fips=1
```

1 enables the FIPS mode. 0, the default, disables the FIPS mode.

Example

```
fips=1
```
maxcpus - Limit the number of CPUs Linux can use at IPL

Use the maxcpus= kernel parameter to limit the number of CPUs that Linux can use at IPL and that are online after IPL.

If the real or virtual hardware provides more than the specified number of CPUs, these surplus CPUs are initially offline. For example, if five CPUs are available, maxcpus=2 results in two online CPUs and three offline CPUs after IPL.

Offline CPUs can be set online dynamically unless the possible_cpus= parameter is set and specifies a maximum number of online CPUs that is already reached. The possible_cpus= parameter sets an absolute limit for the number of CPUs that can be online at any one time (see possible_cpus). If both maxcpus= and possible_cpus= are set, a lower value for possible_cpus= overrides maxcpus= and makes it ineffective.

Format

maxcpus syntax

maxcpus=<number>

Examples

maxcpus=2
nosmt - Disable simultaneous multithreading

By default, Linux in LPAR mode uses simultaneous multithreading if it is supported by the hardware. Specify the \texttt{nosmt} kernel parameter to disable simultaneous multithreading. See also \texttt{smt} - Reduce the number of threads per core\footnote{on page 664}.

For more information about simultaneous multithreading, see \texttt{Simultaneous multithreading}\footnote{on page 327}.

Format

\begin{verbatim}
  nosmt syntax
  >>>nosmt
  <
\end{verbatim}
**possible_cpus**

### possible_cpus - Limit the number of CPUs Linux can use

Use the `possible_cpus=` parameter to set an absolute limit for the number of CPUs that can be online at any one time. If the real or virtual hardware provides more than the specified maximum, the surplus number of CPUs must be offline. Alternatively, you can use the common code kernel parameter `nr_cpus`.

Use the `maxcpus=` parameter to limit the number of CPUs that are online initially after IPL (see `maxcpus`).

#### Format

```plaintext
possible_cpus syntax

possible_cpus=<number>
```

#### Examples

```
possible_cpus=8
```
ramdisk_size - Specify the ramdisk size

Use the `ramdisk_size=` kernel parameter to specify the size of the ramdisk in kilobytes.

**Format**

```
ramdisk_size syntax

ramdisk_size=<size>
```

**Examples**

`ramdisk_size=32000`
ro - Mount the root file system read-only

Use the `ro` kernel parameter to mount the root file system read-only.

**Format**

```
ro syntax
```

```
root - Specify the root device

Use the root= kernel parameter to tell Linux what to use as the root when mounting the root file system.

**Format**

```
root syntax
———root=<rootdevice>———
```

**Examples**

This example makes Linux use /dev/dasda1 when mounting the root file system:

```
root=/dev/dasda1
```
smt - Reduce the number of threads per core

By default, Linux in LPAR mode uses the maximum number of threads per core that is supported by the hardware. Use the smt= kernel parameter to use fewer threads. The value can be any integer in the range 1 to the maximum number of threads that is supported by the hardware.

Specifying smt=1 effectively disables simultaneous multithreading. See also “nosmt - Disable simultaneous multithreading” on page 659.

For more information about simultaneous multithreading, see “Simultaneous multithreading” on page 327.

Format

```
smt syntax

```

where $<hwmax>$ is the maximum number of threads per core that is supported by the hardware, and $<number>$ is an integer in the range 1 - $<hwmax>$.

Examples

```
smt=1
```

vdso - Optimize system call performance

Use the `vdso=` kernel parameter to control the vdso support for the `gettimeofday`, `clock_getres`, and `clock_gettime` system calls.

The virtual dynamic shared object (vdso) support is a shared library that the kernel maps to all dynamically linked programs. The glibc detects the presence of the vdso and uses the functions that are provided in the library.

Because the vdso library is mapped to all user-space processes, this change is visible in user space. In the unlikely event that a user-space program does not work with the vdso support, you can disable the support.

The default, which is to use vdso support, works well for most installations. Do not override this default, unless you observe problems.

The vdso support is included in the Linux kernel.

**Format**

```
vdso syntax

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>on</td>
<td>Enables vdso support</td>
</tr>
<tr>
<td>off</td>
<td>Disables vdso support</td>
</tr>
</tbody>
</table>
```

**Examples**

This example disables the vdso support:

```
vds0=0
```
vmhalt

vmhalt - Specify CP command to run after a system halt

Use the vmhalt= kernel parameter to specify a command to be issued to CP after a system halt.

This command applies only to Linux on z/VM.

Format

\[
\text{vmhalt syntax} \\
\text{\texttt{vmhalt=\textit{COMMAND}}} \\
\]

Examples

This example specifies that an initial program load of CMS is to follow the Linux halt command:

\[
\text{vmhalt="CPU 00 CMD I CMS"} \\
\]

Note: The command must be entered in uppercase.
**vmpanic - Specify CP command to run after a kernel panic**

Use the `vmpanic=` kernel parameter to specify a command to be issued to CP after a kernel panic.

This command applies only to Linux on z/VM.

**Note:** Ensure that the `dumpconf` service is disabled when you use this kernel parameter. Otherwise, `dumpconf` will override the setting.

**Format**

```
vmpanic syntax

vmpanic=<COMMAND>
```

**Examples**

This example specifies that a VMDUMP is to follow a kernel panic:

```
vmpanic="VMDUMP"
```

**Note:** The command must be entered in uppercase.
vmpoff - Specify CP command to run after a power off

Use the vmpoff= kernel parameter to specify a command to be issued to CP after a system power off.

Format

vmpoff syntax

►►vmpoff=<COMMAND>◄◄

Examples

This example specifies that CP is to clear the guest virtual machine after the Linux power off or halt -p command:

vmpoff="SYSTEM CLEAR"

Note: The command must be entered in uppercase.
vmreboot - Specify CP command to run on reboot

Use the `vmreboot=` kernel parameter to specify a command to be issued to CP on reboot.

This command applies only to Linux on z/VM.

Format

```
vmreboot syntax
➤➤vmreboot=<COMMAND>➤◄
```

Examples

This example specifies a message to be sent to the z/VM guest virtual machine OPERATOR if a reboot occurs:

```
vmreboot="MSG OPERATOR Reboot system"
```

Note: The command must be entered in uppercase.
vmreboot
Chapter 55. Linux diagnose code use

Red Hat Enterprise Linux 7.5 for z Systems issues several diagnose instructions to the hypervisor (LPAR or z/VM).

Table 65 lists all diagnoses that are used by the Linux kernel or a kernel module.

Linux can fail if you change the privilege class of the diagnoses marked as **required** by using the MODIFY diag command in z/VM.

**Table 65. Linux diagnoses**

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Linux use</th>
<th>Required/Optional</th>
</tr>
</thead>
</table>
| 0x008  | z/VM CP command console interface | • The **vmcp** command  
• The 3215 and 3270 console drivers  
• The z/VM recording device driver (vmlogrdr)  
• smsgiucv  
• AF_IUCV | Required |
| 0x010  | Release pages | CMM | Required |
| 0x014  | Input spool file manipulation | The vmur device driver | Required |
| 0x044  | Voluntary time-slice end | In the kernel for spinlock and udelay | Required |
| 0x064  | Allows Linux to attach a DCSS | The DCSS block device driver (dcssblk) and the MONITOR record device driver (monreader). | Required |
| 0x09c  | Voluntary time slice yield | Spinlock. | Optional |
| 0x0c   | z/VM management time data | The hyptop command. | Optional |
| 0x0dc  | Monitor stream | The APPLDATA monitor record and the MONITOR stream application support (monwriter). | Required |
| 0x204  | LPAR Hypervisor data | The hypervisor file system (hypfs). | Required |
| 0x210  | Retrieve device information | • The common I/O layer  
• The DASD driver DIAG access method  
• DASD read-only query  
• The vmur device driver | Required |
| 0x224  | CPU type name table | The hypervisor file system (hypfs). | Required |
| 0x250  | Block I/O | The DASD driver DIAG access method. | Required |
| 0x258  | Page-reference services | In the kernel, for pfault. | Optional |
| 0x26c  | z/VM: Access Certain System Information | The qeth device driver, for MAC address | Optional |
| 0x288  | Virtual machine time bomb | The watchdog device driver. | Required |
Table 65. Linux diagnoses (continued)

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
<th>Linux use</th>
<th>Required/Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x2fc</td>
<td>Hypervisor cpu and memory accounting data</td>
<td>The hypervisor file system (hypfs).</td>
<td>Required</td>
</tr>
<tr>
<td>0x308</td>
<td>Re-ipl</td>
<td>Re-ipl and dump code.</td>
<td>Required</td>
</tr>
<tr>
<td>0x500</td>
<td>Virtio functions</td>
<td>Operate virtio-ccw devices</td>
<td>Required</td>
</tr>
</tbody>
</table>

Required means that a function is not available without the diagnose; optional means that the function is available but there might be a performance impact.
Part 11. Appendixes
Appendix A. Accessibility

Accessibility features help users who have a disability, such as restricted mobility or limited vision, to use information technology products successfully.

Documentation accessibility

The Linux on Z and LinuxONE publications are in Adobe Portable Document Format (PDF) and should be compliant with accessibility standards. If you experience difficulties when you use the PDF file and want to request a Web-based format for this publication, use the Readers’ Comments form in the back of this publication, send an email to eservdoc@de.ibm.com, or write to:

IBM Deutschland Research & Development GmbH
Information Development
Department 3282
Schoenaicher Strasse 220
71032 Boeblingen
Germany

In the request, be sure to include the publication number and title.

When you send information to IBM, you grant IBM a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

IBM and accessibility

See the IBM Human Ability and Accessibility Center for more information about the commitment that IBM has to accessibility at

www.ibm.com/able
Appendix B. Understanding syntax diagrams

This section describes how to read the syntax diagrams in this manual.

To read a syntax diagram follow the path of the line. Read from left to right and top to bottom.

- The ►► symbol indicates the beginning of a syntax diagram.
- The ───► symbol, at the end of a line, indicates that the syntax diagram continues on the next line.
- The ►── symbol, at the beginning of a line, indicates that a syntax diagram continues from the previous line.
- The ───►◄ symbol indicates the end of a syntax diagram.

Syntax items (for example, a keyword or variable) may be:

- Directly on the line (required)
- Above the line (default)
- Below the line (optional)

If defaults are determined by your system status or settings, they are not shown in the diagram. Instead the rule is described together with the option, keyword, or variable in the list following the diagram.

**Case sensitivity**

Unless otherwise noted, entries are case sensitive.

**Symbols**

You must code these symbols exactly as they appear in the syntax diagram

- Asterisk
- Colon
- Comma
- Equal sign
- Hyphen
- Double slash
- Parentheses
- Period
- Add
- Dollar sign

For example:

dasd=0.0.7000-0.0.7fff

**Variables**

An italicized lowercase word enclosed in angled brackets indicates a variable that you must substitute with specific information. For example:

```
►► -p ◄interface►◄
```

Here you must code -p as shown and supply a value for <interface>.
An italicized uppercase word in angled brackets indicates a variable that must appear in uppercase:

►►vmhalt◄◄<COMMAND>◄◄

**Repetition**

An arrow returning to the left means that the item can be repeated.

►►<repeat>◄◄

A character within the arrow means you must separate repeated items with that character.

►►,◄◄<repeat>◄◄

**Defaults**

Defaults are above the line. The system uses the default unless you override it. You can override the default by coding an option from the stack below the line. For example:

►►A
   B
   C◄◄

In this example, A is the default. You can override A by choosing B or C.

**Required Choices**

When two or more items are in a stack and one of them is on the line, you must specify one item. For example:

►►A
   B
   C◄◄

Here you must enter either A or B or C.

**Optional Choice**

When an item is below the line, the item is optional. Only one item may be chosen. For example:

►►A
   B
   C◄◄

Here you may enter either A or B or C, or you may omit the field.
Notices

This information was developed for products and services offered in the U.S.A. IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

The licensed program described in this information and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement, or any equivalent agreement between us.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.
This information is for planning purposes only. The information herein is subject to change before the products described become available.

**Trademarks**

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at [www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml)

Adobe is either a registered trademark or trademark of Adobe Systems Incorporated in the United States, and/or other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.
Glossary

This glossary includes IBM product terminology as well as selected other terms and definitions.

Additional information can be obtained in:


- The Information Technology Vocabulary developed by Subcommittee 1, Joint Technical Committee 1, of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC JTC1/SC1).


- Internet Request for Comments: 1208, Glossary of Networking Terms

- Internet Request for Comments: 1392, Internet Users’ Glossary


A

**address space.** The range of addresses available to a computer program or process. Address space can refer to physical storage, virtual storage, or both.

**auto-detection.** Listing the addresses of devices attached to a card by issuing a query command to the card.

C

**CEC.** (Central Electronics Complex). A synonym for CPC.

**channel subsystem.** The programmable input/output processors of the mainframe, which operate in parallel with the CPU.

**checksum.** An error detection method using a check byte appended to message data.

**CHPID.** channel path identifier. In a channel subsystem, a value assigned to each installed channel path of the system that uniquely identifies that path to the system.

**compatible disk layout.** A disk structure for Linux on Z which allows access from other mainframe operating systems. This replaces the older Linux disk layout.

**Console.** In Linux, an output device for kernel messages.

**CPC.** (Central Processor Complex). A physical collection of hardware that includes main storage, one or more central processors, timers, and channels. Also referred to as a CEC.

**CRC.** cyclic redundancy check. A system of error checking performed at both the sending and receiving station after a block-check character has been accumulated.

**CSMA/CD.** carrier sense multiple access with collision detection.

**CTC.** channel to channel. A method of connecting two computing devices.

**CUU.** control unit and unit address. A form of addressing for mainframe devices using device numbers.
D

DASD. direct access storage device. A mass storage medium on which a computer stores data.

device driver.
• A file that contains the code needed to use an attached device.
• A program that enables a computer to communicate with a specific peripheral device; for example, a printer, a videodisc player, or a CD-ROM drive.
• A collection of subroutines that control the interface between I/O device adapters and the processor.

DIAGNOSE. In z/VM, a set of instructions that programs running on z/VM guest virtual machines can call to request CP services.

disconnected device. In Linux on Z, a device that is online, but to which Linux can no longer find a connection. Reasons include:
• The device was physically removed
• The device was logically removed, for example, with a CP DETACH command in z/VM
• The device was varied offline

E

ECKD. extended count-key-data device. A disk storage device that has a data transfer rate faster than some processors can utilize and that is connected to the processor through use of a speed matching buffer. A specialized channel program is needed to communicate with such a device.

ESCON. enterprise systems connection. A set of IBM products and services that provide a dynamically connected environment within an enterprise.

Ethernet. A 10-Mbps baseband local area network that allows multiple stations to access the transmission medium at will without prior coordination, avoids contention by using carrier sense and deference, and resolves contention by using collision detection and delayed retransmission. Ethernet uses CSMA/CD.

F

Fast Ethernet (FENET). Ethernet network with a bandwidth of 100 Mbps

FBA. fixed block architecture. An architecture for a virtual device that specifies the format of and access mechanisms for the virtual data units on the device. The virtual data unit is a block. All blocks on the device are the same size (fixed size). The system can access them independently.


fibre channel. A technology for transmitting data between computer devices. It is especially suited for attaching computer servers to shared storage devices and for interconnecting storage controllers and drives.

FTP. file transfer protocol. In the Internet suite of protocols, an application layer protocol that uses TCP and Telnet services to transfer bulk-data files between machines or hosts.

G

Gigabit Ethernet (GbE). An Ethernet network with a bandwidth of 1000-Mbps

H

hardware console. A service-call logical processor that is the communication feature between the main processor and the service processor.

Host Bus Adapter (HBA). An I/O controller that connects an external bus, such as a Fibre Channel, to the internal bus (channel subsystem).

In a Linux environment HBAs are normally virtual and are shown as an FCP device.

HMC. hardware management console. A console used to monitor and control hardware such as the IBM Z microprocessors.

HFS. hierarchical file system. A system of arranging files into a tree structure of directories.

I

intraensemble data network (IEDN). A private 10 Gigabit Ethernet network for application data communications within an ensemble. Data communications for workloads can flow over the IEDN within and between nodes of an ensemble. All of the physical and logical resources of the IEDN are configured, provisioned, and managed by the Unified Resource Manager.

intranode management network (INMN). A private 1000BASE-T Ethernet network operating at 1 Gbps that is required for the Unified Resource Manager to manage the resources within a single zEnterprise node. The INMN connects the Support Element (SE) to the zEnterprise 196 (z196) or zEnterprise 114 (z114) and to any attached zEnterprise BladeCenter Extension (zBX).

ioctl system call. Performs low-level input- and output-control operations and retrieves device status
information. Typical operations include buffer manipulation and query of device mode or status.

**IOCS.** input / output channel subsystem. See channel subsystem.

**IP.** internet protocol. In the Internet suite of protocols, a connectionless protocol that routes data through a network or interconnected networks and acts as an intermediary between the higher protocol layers and the physical network.

**IP address.** The unique 32-bit address that specifies the location of each device or workstation on the Internet. For example, 9.67.97.103 is an IP address.

**IPIP.** IPv4 in IPv4 tunnel, used to transport IPv4 packets in other IPv4 packets.

**IPL.** initial program load (or boot).

- The initialization procedure that causes an operating system to commence operation.
- The process by which a configuration image is loaded into storage at the beginning of a work day or after a system malfunction.
- The process of loading system programs and preparing a system to run jobs.


**IUCV.** inter-user communication vehicle. A z/VM facility for passing data between virtual machines and z/VM components.

---

**K**

**kernel.** The part of an operating system that performs basic functions such as allocating hardware resources.

**kernel module.** A dynamically loadable part of the kernel, such as a device driver or a file system.

**kernel image.** The kernel when loaded into memory.

---

**L**

**LCS.** LAN channel station. A protocol used by OSA.

**LDP.** Linux Documentation Project. An attempt to provide a centralized location containing the source material for all open source Linux documentation. Includes user and reference guides, HOW TOs, and FAQs. The homepage of the Linux Documentation Project is [www.linuxdocs.org](http://www.linuxdocs.org).

**Linux.** a variant of UNIX which runs on a wide range of machines from wristwatches through personal and small business machines to enterprise systems.

**Linux disk layout.** A basic disk structure for Linux on Z. Now replaced by compatible disk layout.

**Linux on Z.** the port of Linux to the IBM mainframe architecture.

**LPAR.** logical partition of a mainframe. Logical partitions are, in practice, equivalent to separate mainframes; each LPAR runs its own operating system.

**LVS (Linux virtual server).** Network sprayer software used to dispatch, for example, http requests to a set of web servers to balance system load.

---

**M**

**MAC.** medium access control. In a LAN this is the sub-layer of the data link control layer that supports medium-dependent functions and uses the services of the physical layer to provide services to the logical link control (LLC) sub-layer. The MAC sub-layer includes the method of determining when a device has access to the transmission medium.

**Mbps.** million bits per second.

**MIB (Management Information Base).**

- A collection of objects that can be accessed by means of a network management protocol.
- A definition for management information that specifies the information available from a host or gateway and the operations allowed.

**MTU.** maximum transmission unit. The largest block which may be transmitted as a single unit.

**Multicast.** A protocol for the simultaneous distribution of data to a number of recipients, for example live video transmissions.

---

**N**

**NIC.** network interface card. The physical interface between the IBM mainframe and the network.

---

**O**

**OSA-Express.** Abbreviation for Open Systems Adapter-Express networking features. These include 10 Gigabit Ethernet, Gigabit Ethernet, and Fast Ethernet.

**OSM.** OSA-Express for Unified Resource Manager. A CHPID type that provides connectivity to the intranode management network (INMN) from z196 or z114 to Unified Resource Manager functions. Uses OSA-Express3 1000BASE-T Ethernet exclusively operating at 1 Gbps.

**OSPF.** open shortest path first. A function used in route optimization in networks.
OSX. OSA-Express for zBX. A CHPID type that provides connectivity and access control to the intraensemble data network (IEDN) from z196 or z114 to zBX.

POR. power-on reset

POSIX. Portable Operating System Interface for Computer Environments. An IEEE operating system standard closely related to the UNIX system.

router. A device or process which allows messages to pass between different networks.

SE. support element.

- An internal control element of a processor that assists in many of the processor operational functions.
- A hardware unit that provides communications, monitoring, and diagnostic functions to a central processor complex.

SNA. systems network architecture. The IBM architecture that defines the logical structure, formats, protocols, and operational sequences for transmitting information units through, and controlling the configuration and operation of, networks. The layered structure of SNA allows the ultimate origins and destinations of information (the users) to be independent of and unaffected by the specific SNA network services and facilities that are used for information exchange.

SNMP (Simple Network Management Protocol). In the Internet suite of protocols, a network management protocol that is used to monitor routers and attached networks. SNMP is an application layer protocol. Information about devices managed is defined and stored in the application's Management Information Base (MIB).

Sysctl. system control programming manual control (frame). A means of dynamically changing certain Linux kernel parameters during operation.

V=V. In z/VM, a guest whose real memory (virtual from a z/VM perspective) corresponds to virtual memory of z/VM.

V=V. In z/VM, a guest whose real memory (virtual from a z/VM perspective) corresponds to virtual memory of z/VM.

Virtual LAN (VLAN). A group of devices on one or more LANs that are configured (using management software) so that they can communicate as if they were attached to the same wire, when in fact they are located on a number of different LAN segments. Because VLANs are based on logical rather than physical connections, they are extremely flexible.

volume. A data carrier that is usually mounted and demounted as a unit, for example a tape cartridge or a disk pack. If a storage unit has no demountable packs the volume is the portion available to a single read/write mechanism.

z114. IBM zEnterprise 114

z13. IBM z13

z14. IBM z14.

z13s. IBM z13s.

z196. IBM zEnterprise 196

zBC12. IBM zEnterprise BC12.

zBX. IBM zEnterprise BladeCenter Extension
zEC12. IBM zEnterprise EC 12.

zEnterprise. IBM zEnterprise System. A heterogeneous hardware infrastructure that can consist of an IBM zEnterprise BC12, a zEnterprise EC12 (zEC12), a zEnterprise 114 (z114) or a zEnterprise 196 (z196) and an attached IBM zEnterprise BladeCenter Extension (zBX), managed as a single logical virtualized system by the Unified Resource Manager.

zSeries. The family of IBM enterprise servers that demonstrate outstanding reliability, availability, scalability, security, and capacity in today's network computing environments.
Bibliography

The publications listed here are considered useful for a more detailed study of the topics contained in this book.

Linux on Z and LinuxONE publications

The Linux on Z and LinuxONE publications can be found on the developerWorks website.

You can find the latest versions of this publication at


- Using the Dump Tools on Red Hat Enterprise Linux 7, SC34-2711

For each of the following publications, the same web page points to the version that most closely reflects Red Hat Enterprise Linux 7.5:

- How to Improve Performance with PAV, SC33-8414
- How to use FC-attached SCSI devices with Linux on z Systems, SC33-8413
- How to Set up a Terminal Server Environment on z/VM, SC34-2596
- Exploiting Enterprise PKCS #11 using openCryptoki, SC34-2713
- libica Programmer’s Reference, SC34-2602
- Secure Key Solution with the Common Cryptographic Architecture Application Programmer’s Guide, SC33-8294
- Linux on z Systems Troubleshooting, SC34-2612

Red Hat Enterprise Linux 7.5 publications

The documentation for Red Hat Enterprise Linux 7.5 can be found on the Red Hat documentation website.

See

https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux

- Red Hat Enterprise Linux 7.5 Installation Guide
- Red Hat Enterprise Linux 7.5 System Administrator’s Guide
- Red Hat Enterprise Linux 7.5 Networking Guide

z/VM publications

The publication numbers listed are for z/VM version 6.

For the complete library including other versions, see

www.ibm.com/vm/library

- z/VM Connectivity, SC24-6174
- z/VM CP Commands and Utilities Reference, SC24-6175
- z/VM CP Planning and Administration, SC24-6178
- z/VM CP Programming Services, SC24-6179
- z/VM Getting Started with Linux on System z, SC24-6194
- z/VM Performance, SC24-6208
IBM Redbooks publications

You can search for, view, or download Redbooks publications, Redpapers™, Hints and Tips, draft publications and additional materials on the Redbooks website.

You can also order hardcopy Redbooks or CD-ROMs. See www.ibm.com/redbooks

- IBM zEnterprise Unified Resource Manager, SG24-7921
- Building Linux Systems under IBM VM, REDP-0120
- FICON CTC Implementation, REDP-0158
- Networking Overview for Linux on zSeries, REDP-3901
- Linux on IBM eServer zSeries and S/390: TCP/IP Broadcast on z/VM Guest LAN, REDP-3596
- Linux on IBM eServer zSeries and S/390: VSWITCH and VLAN Features of z/VM 4.4, REDP-3719
- Security on z/VM, SG24-7471
- IBM Communication Controller Migration Guide, SG24-6298
- Problem Determination for Linux on System z, SG24-7599
- Fibre Channel Protocol for Linux and z/VM on IBM System z, SG24-7266

Other IBM Z publications

General Z publications that might be of interest in the context of Linux on Z.

- zEnterprise System Introduction to Ensembles, GC27-2609
- zEnterprise System Ensemble Planning and Configuring Guide, GC27-2608
- System z Application Programming Interfaces, SB10-7030
- IBM TotalStorage Enterprise Storage Server System/390 Command Reference 2105 Models E10, E20, F10, and F20, SC26-7295
- Processor Resource/Systems Manager Planning Guide, SB10-7041
- z/Architecture Principles of Operation, SA22-7832
- z/Architecture The Load-Program-Parameter and the CPU-Measurement Facilities, SA23-2260
- IBM The CPU-Measurement Facility Extended Counters Definition for z10, z196, z114 and zEC12, SA23-2261

Networking publications

- HiperSockets Implementation Guide, SG24-6816
- Open Systems Adapter-Express Customer’s Guide and Reference, SA22-7935
- OSA-Express Implementation Guide, SG24-5948

Security related publications

- zSeries Crypto Guide Update, SG24-6870
ibm.com resources

On the ibm.com® website you can find information about many aspects of Linux on Z and IBM LinuxONE™ including z/VM, I/O connectivity, and cryptography.

- For CMS and CP Data Areas, Control Block information, and the layout of the z/VM monitor records see
- For I/O connectivity information, see
  www.ibm.com/systems/z/connectivity
- For Communications server for Linux information, see
  www.ibm.com/software/network/commserver/linux
- For information about performance monitoring on z/VM, see
  www.ibm.com/vm/perf
- For cryptographic coprocessor information, see
  www.ibm.com/security/cryptocards
- (Requires registration.) For information for planning, installing, and maintaining IBM systems, see
  www.ibm.com/servers/resourcelink
- For information about STP, see
  www.ibm.com/systems/z/advantages/pso/stp.html
Index

Special characters

/debug, mount point  xi
/proc, mount point  xi
/sys, mount point  xi
/sys/devices  7
/sys/kernel/debug, mount point  xi
*ACCOUNT, z/VM record  391
*LOGREC, z/VM record  391
*SYMPTOM, z/VM record  391

Numerics

10 Gigabit Ethernet
SNMP  289
1000Base-T Ethernet
LAN channel station  299
SNMP  289
1750, control unit  121
2015, control unit  121
2107, control unit  121
3088, control unit  121
3091, control unit  121
3270 emulation  45
3270 terminal device driver  44
switching the views of  47
3370, DASD  121
3380, DASD  121
3390, DASD  121
3392, control unit  121
3990, control unit  121
3DES  441
6310, control unit  121
9336, DASD  121
9343, control unit  121
9345, DASD  121

A
acceleration, in-kernel cryptography  449
access control
  osasnmtp  291
access_denied
  zfcp attribute (port)  178
  zfcp attribute (SCSI device)  185
access_shared
  zfcp attribute  185
accessibility  675
ACCOUNT, z/VM record  391
actions, shutdown  109
adapter outage  273
adapter virtualization  428
add, DCSS attribute  405
adding and removing cryptographic adapters  436
Address Resolution Protocol
  See ARP
AES  449
aes_s390, kernel module  450
AF_IUCV
  addressing sockets in applications  321
  set up devices for addressing  320
AF_IUCV address family  319
  features  319
  set up support for  320
af_iucv, kernel module  321
AgentX protocol  289
alias
  DASD attribute  156
ap
  module parameter  430
AP
  devices  7
  AP bus
    attributes  438
AP queue  429
ap_functions
  cryptographic adapter attribute  432
ap_interrupt
  cryptographic adapter attribute  435
API
  cryptographic  439
  FC-HBA  163
  GenWQE zlib  367
  zfcp HBA  196
APPDATA monitor records  371
  monitoring Linux instances  371
APPDATA, monitor stream  375
applet
  emulation of the HMC Operating System Messages  51
applications
  addressing AF_IUCV sockets in  321
arch
  trng counter  446
ARP  235
  proxy ARP  267
  query/purge OSA-Express ARP cache  608
attributes
  device  9
  for CCW devices  9
  for subchannels  13
qeth  239, 240
setting  10
authorization
  CPU-measurement counter facility  466
auto-detection
  DASD  132
autoconfiguration, IPv6  231
automatic problem reporting
  activating  479
autopurge, z/VM recording attribute  394
autorecording, z/VM recording attribute  393
availability
  common CCW attribute  9
  DASD attribute  137
avg_*, cmf attributes  459
avg_control_unit_queuing_time, cmf attribute  459
avg_device_active_only_time, cmf attribute  459
avg_device_busy_time  459
avg_device_busy_time, cmf attribute  459

© Copyright IBM Corp. 2000, 2018

691
avg_device_connect_time, cmf attribute 459
avg_device_disconnect_time, cmf attribute 459
avg_function_pending_time, cmf attribute 459
avg_initial_command_response_time, cmf attribute 459
avg_sample_interval, cmf attribute 459
avg_utilization, cmf attribute 459
B
base device
  helper script 65
block_size_bytes, memory attribute 335
blocksize, tape attribute 211
book_siblings
  CPU sysfs attribute 330
boot configuration
  module parameters 29
boot devices 82
  logical 62
  preparing 55
boot loader code 83
boot menu
  DASD, z/VM example 86
  zipl 71
booting Linux 81
troubleshooting 485
bridge_hostnotify, qeth attribute 235
bridge_role, qeth attribute 235, 269
bridge_state, qeth attribute 235
buffer_count, qeth attribute 246
buffer-overflow protection 453
buffer, CPU-measurement sampling facility 468
buffer, CTCM attribute 311
bus ID 8
byte_counter
  prandom attribute 443
  trng attribute 446
C
cachesize=, module parameters 356
Call Home
  callhome attribute 479
callhome
    Call Home attribute 479
capability change, CPU 328
card_type, qeth attribute 247
card_version, zfcp attribute 169
case conversion 52
CBC 449
CCW
  channel measurement facility 457
  common attributes 9
devices 7
group devices 7
hotplug events 18
setting attributes 490
setting devices online/offline 490
CCW terminal device
  switching on- or offline 48
CD-ROM, loading Linux 95
Central Processor Assist for Cryptographic Function
  See CPACF
CEX3A (Crypto Express3) 425
CEX3C (Crypto Express3) 425
CEX4A (Crypto Express4S) 425
CEX4C (Crypto Express4S) 425
CEX4P (Crypto Express4S) 425
CEX5A (Crypto Express5S) 425
CEX5C (Crypto Express5S) 425
CEX5P (Crypto Express5S) 425
change, CPU capability 328
channel measurement facility 457
  cmb_enable attribute 458
  features 457
  kernel parameters 457
  read-only attributes 458
channel path
  changing status 492
  determining usage 484
  ensuring correct status 483
  list 562
channel path availability
  planned changes 483
  unplanned changes 483
channel path ID 15
channel path measurement 14
channel subsystem view 12
channel-attached tape 205
chccwdev 10
chccwdev, Linux command 490
chcp, Linux command 492
chcpu, Linux command 327
chcpumpf, Linux command 496
checksum
  inbound 253
  outbound 253
CHID
  mapping physical to virtual 17
Chinese-Remainder Theorem 425
chiuscvallow, Linux command 42
chmem, Linux command 494
CHPID
  in sysfs 15
  map to PCHID 17
  online attribute 15, 16
  chipids, subchannel attribute 14
  chreipl, Linux command 497
  chshut, Linux command 501
  chunksize
    prandom attribute 443
    trng attribute 446
    cmm
    avoid swapping with 373
    background information 373

692 Device Drivers, Features, and Commands - Red Hat Enterprise Linux 7.5
CMM
unload module 484

cmm, kernel module 421

CMMA 656

cmma=, kernel parameter 656

CMS disk layout 126

CMS1 labeled disk 126

cmsfs-fuse, Linux command 505

code page 506
for x3270 45

Collaborative Memory Management Assist 656

collecting QETH performance statistics 257

command
qetharp 608

commands, Linux
chccwdev 490
chchp 492
chcp 327
chcvmf 496
chiucvallow 42
chshut 501
chzcrypt 503

cmsfs-fuse 505

cpuplugd 510, 511

dasdstat 525

dasdview 528

dmesg 5

dumpconf 110

execstack 453

fdasd 539

genwqe_echo 362

genwqe_gunzip 362

genwqe_gzip 362

gunzip 364

gzip 364

hmcdrvs 547

yhtop 551

icainfo 489

castats 489

ip 4

iucvconn 43

iucvtty 43

lchp 562

lsCPU 327

lscpu 327

lscpumf 564

lsCSS 567

lsdas 571

lshm 575

lsLuns 576

lsmem 579

lsqeth 581

lsreip1 583

lsscm 584

lsshut 586

lstape 587

lszcrypt 591

lszfcp 595

mon_fsstatd 597

mon_proc 601

qetharp 608

qethconf 610

qethqot 613

readelf 453

readlink 5

commands, Linux (continued)

scsi_logging_level 616

time 364

tunedasd 625

vmconvert 636

vmcp 629

vmur 631

zdsfs 640

zfcp_ping 198

zfcp_show 198

zipl 55

znetconf 648

commands, z/VM

sending from Linux 629

communication facility

Inter-User Communication Vehicle 319

compatible disk layout 123

compression

GenWQE 359

compression, tape 212

conceal=, module parameters 116

CONFIG_FUSE_FS 640

configuration file

CPU control 512

cpuplugd 518

memory control 513

configure LPAR I/O devices 484

configuring standby CPU 328

conmode=, kernel parameter 41

console

definition 35

device names 35

device nodes 36

mainframe versus Linux 35

console device driver

kernel parameter 41

overriding default driver 41

restricting access to HVC terminal devices 42

SCLP line-mode buffer page reuse 42

SCLP line-mode buffer pages 42

specifying preferred console 41

specifying the number of HVC terminal devices 42

console device drivers 33

device and console names 35

features 34

terminal modes 36

console=, kernel parameter 41

control characters 49

control program identification 475

control unit

1750 121

2105 121

2107 121

3880 121

3990 121

6310 121

9343 121

controlling automatic port scans 175

cooperative memory management 421

set up 421

core 327

core_siblings

CPU sysfs attribute 330
DASD (continued)
forcing online 137
formatting ECKD 520
High Performance FICON 148
hpf attribute 154
last_known_reservation_state attribute 152
Linux disk layout 126
module parameter 131
online attribute 140, 141
partitioning 539, 551
partitions on 123
path_interval attribute 153
path_threshold attribute 153
PAV 148
performance statistics 525
performance tuning 625
raw_track_access attribute 149
readonly attribute 156
reservation_policy attribute 149
statistics 144
status attribute 157
uid attribute 157
vendor attribute 157
virtual 121
volume label 124
DASD information
  displaying 155
dasd=
  module parameter 131
dasdfmt, Linux command 520
dasdstat, Linux command 525
dasdview, Linux command 528
data
  compression 359
data consistency checking, SCSI 194
data integrity extension 194
data integrity field 194
datarouter=
  zfcp module parameter 165
dbfs=, zfcp module parameters 165
DCSS 387
  access mode 406
  add attribute 405
  adding 405
  device driver 401
  device names 401
  device nodes 401
  exclusive-writable mode 401
  minor number 406
  performance monitoring using 372
  remove attribute 409
  save attribute 408
  saving with properties 408
  seglist attribute 405
  shared attribute 407
  with options 402
dcsblk.segments=, module parameter 402
deactivating a qeth interface 251
debug feature 374
decompression, GenWQE 359
decryption 425
default hugepagesize=, kernel parameters 339
delete
  zfcp sysfs attribute 192
  delete, zfcp attribute 193
depth
cryptographic adapter attribute 432
des_s390, kernel module 450
determine channel path usage 484
device bus-ID 8
  of a qeth interface 249
device driver
cryptographic 425
CTCM 305
DASD 121
DCSS 401
Generic Work Queue Engine 359
HiperSockets 223
HMC media 355
LCS 299
mlb4_en 323
monitor stream application 381
network 219
OSA-Express (QDIO) 223
overview 8
PCIe 19
protected key 447
pseudorandom number 441
qeth 223
SCLP_ASYNC 479
SCSI-over-Fibre Channel 159
smgguv_app 415
storage-class memory 201
tape 205
ture random number 445
vmcp 413
vmur 399
watchdog 115
XPRAM 215
z/VM *MONITOR record reader 385
z/VM recording 391
zcrypt 425
device drivers
  support of the FCP environment 160
device names 3
  console 35
  DASD 128
  DCSS 401
  storage class memory 201
tape 206
  vmur 399
  watchdog 115
  XPRAM 215
  z/VM *MONITOR record 386
  z/VM recording 391
device node
  random, non-root users 442
  z90crypt 431
device nodes 3
  console 36
  DASD 128
  DCSS 401
  GenWQE 361
  SCSI 161
  storage class memory 201
tape 207
  vmcp 413
  vmur 399
  watchdog 115
  z/VM *MONITOR record 386
  z/VM recording 391
  zfcp 161
device numbers 3
device special file
   See device nodes
device view 12
   by category 11
   by device drivers 11
device_blocked
   zfcp attribute (SCSI device) 186
devices
   alias 156
   attributes 9
   base 156
   corresponding interfaces 5
   ignoring 652
   in sysfs 8
   working with newly available 10
devs=. module parameter 216
devtype
   common CCW attribute 9
   tape attribute 211
dhcp 286
DHCP 285
   required options 285
DIAG
   access method 138
   DIAG access method 138
   for ECKD 127
   for FBA 127
DIAG call 671
diag288 watchdog
device driver 115
diag288_wdt
   kernel module 116
diagnose call 671
diagnosis
   using XPRAM 216
DIF 194
difs
   zfcp module parameter 165
Direct Access Storage Device
   See DASD
Direct SNMP 289
disabled wait
   booting stops with 485
cio_ignore 484
discipline
   DASD attribute 156
   discontiguous saved segments
   See DCSS
disk layout
   CMS 126
   LDL 126
   summary 127
   Z compatible 123
dispatching
   CPU sysfs attribute 331
displaying
   DASD information 155
displaying information
   FCP channel and device 169
DIX 194
dmesg 5
Documentation directory  xi
documentation website
   Red Hat Enterprise Linux 687
domain, cryptographic 428
domain=
   module parameter 430
dracut 29
drawer_siblings
   CPU sysfs attribute 330
drivers
   See device driver
dsn
   metadata driver attribute 640
dsson
   metadata file attribute 640
dump
   creating automatically after kernel panic 485
dump device
   DASD and tape 66
   ECKD DASD 67
   SCSI 68
dump file
   receive and convert 637
dumpconf, Linux command 110
dumped_frames, zfcp attribute 170
DVD drive, HMC 355
DVD, loading Linux 95
Dynamic Host Configuration Protocol
   See DHCP
dynamic routing, and VIPA 272

E
EADM subchannels
   list 202
EBCDIC 23
   conversion through cmsfs-fuse 505
   kernel parameters 83
ECB 449
ECKD 121
   devices 121
   disk layout summary 127
   raw_track_access attribute 149
ECKD type DASD 135
   preparing for use 135
   edit characters, z/VM console 54
EEDK 619
eer_enabled
   DASD attribute 139
EKM 619
   emulation of the HMC Operating System Messages applet 51
   enable, qeth IP takeover attribute 264
   encoding 506
   encryption 425
   encryption key manager 619
   end-of-line character 53
   end-to-end data consistency, SCSI 194
Enterprise Storage Server 121
   environment variable 417
   environment variables 417
   for CP special messages 417
   TERM 43
   ZIPCONF 76
   ZLIB_CARD 362
   ZLIB_DEFLATE_IMPL 362
   ZLIB_INFLATE_IMPL 362
   ZLIB_TRACE 362
EP11 425
ep11.h 439
erplog, DASD attribute 142
Error Logging System Service
   417
   error_frames, zfcp attribute 391

696  Device Drivers, Features, and Commands - Red Hat Enterprise Linux 7.5
errorflag
prandom attribute 443
escape character
for terminals 53
ESS 121
Ethernet
interface name 231
LAN channel station 299
etr
online attribute 351
ETR 349, 351
kernel parameter 350
etr=, kernel parameter 350
exclusive-writable mode
DCSS access 401
execstack, Linux command 453
expanded memory 215
expires, DASD attribute 143
extended error reporting
DASD 139
extended error reporting, DASD 122
extended remote copy 349
external encrypted data key 619
external time reference 349

F
failed
zcpi attribute (channel) 172
zcpi attribute (port) 179
failfast, DASD attribute 142
fake_bcast, qeth attribute 263
Fast Ethernet
LAN channel station 299
FBA
disk layout summary 127
FBA devices 121
FBA type DASD
preparing for use 136
FC-HBA 163
FC-HBA API functions 197
FCP 159
channel 159
debbuging 165
device 159
trances 165
FCP channel
displaying information 169
FCP device
displaying information 169
FCP devices
listing 195
status information 174
sysfs structure 160
FCP environment 160
fcp_control_requests zcpi attribute 170
fcp_input_megabytes zcpi attribute 170
fcp_input_requests zcpi attribute 170
fcp_lun
zcpi attribute (SCSI device) 186
fcp_lun, zcpi attribute 184
fcp_output_megabytes zcpi attribute 170
fcp_output_requests zcpi attribute 170
fdasd
menu example 543
options, example 546

information
zcpi attribute (channel) 172
zcpi attribute (port) 179
failfast, DASD attribute 142
fake_bcast, qeth attribute 263
Fast Ethernet
LAN channel station 299
FBA
disk layout summary 127
FBA devices 121
FBA type DASD
preparing for use 136
FC-HBA 163
FC-HBA API functions 197
FCP 159
channel 159
debbuging 165
device 159
trances 165
FCP channel
displaying information 169
FCP device
displaying information 169
FCP devices
listing 195
status information 174
sysfs structure 160
FCP environment 160
fcp_control_requests zcpi attribute 170
fcp_input_megabytes zcpi attribute 170
fcp_input_requests zcpi attribute 170
fcp_lun
zcpi attribute (SCSI device) 186
fcp_lun, zcpi attribute 184
fcp_output_megabytes zcpi attribute 170
fcp_output_requests zcpi attribute 170
fdasd
menu example 543
options, example 546

F
failed
zcpi attribute (channel) 172
zcpi attribute (port) 179
failfast, DASD attribute 142
fake_bcast, qeth attribute 263
Fast Ethernet
LAN channel station 299
FBA
disk layout summary 127
FBA devices 121
FBA type DASD
preparing for use 136
FC-HBA 163
FC-HBA API functions 197
FCP 159
channel 159
debbuging 165
device 159
trances 165
FCP channel
displaying information 169
FCP device
displaying information 169
FCP devices
listing 195
status information 174
sysfs structure 160
FCP environment 160
fcp_control_requests zcpi attribute 170
fcp_input_megabytes zcpi attribute 170
fcp_input_requests zcpi attribute 170
fcp_lun
zcpi attribute (SCSI device) 186
fcp_lun, zcpi attribute 184
fcp_output_megabytes zcpi attribute 170
fcp_output_requests zcpi attribute 170
fdasd
menu example 543
options, example 546

INDEX
54
547
547
547
547
547
547
547
547
547
547
547
547
hardware
  random number 445
  service level 485
hardware counter
  reading with perf tool 466
hardware facilities 455
Hardware Management Console
  See HMC
hardware sampling
  set up LPAR 461
hardware status, cryptographic adapter
  online 433
hardware_version, zfcp attribute 169
hardware-acceleration, in-kernel cryptography 449
HBA API 163
  developing applications that use 196
  running applications that use 198
HBA API support
  zfcp 196
hba_id
  zfcp attribute (SCSI device) 186
  zfcp attribute 184
High Performance FICON 148
High Performance FICON, suppressing 132
high resolution polling timer 503
HiperSockets
  bridge port 235
  device driver 223
  interface name 231
  network traffic analyzer 286
HiperSockets Network Concentrator 280
historical data
  cpuplugd keywords 517
HMC 33
  as terminal 46
  definition 35
  for booting Linux 82
  Integrated ASCII console applet 37, 38
  Operating System Messages applet 37
  using in LPAR 37
  using on z/VM 38
HMC DVD drive 357, 358
HMC media
  list media contents 575
  mount media 547
HMC media, device driver 355
HMC Operating System Messages applet
  emulation of the 51
HMC removable media
  assign to LPAR 357
hmc_network attribute 353
hmcdrv.cachesize=, kernel parameters 355
hmcdrvfs, kernel module 356
hmcdrvfs, Linux command 547
hotplug
  CCW devices 18
  memory 333
hotplug memory
  in sysfs 333
  large pages 340
  reboot 334
hotplug rules
  CPU 515
  memory 516

hpf
  DASD attribute 154
hsuid, qeth attribute 269
hugepages=, kernel parameters 339
hugetlbfs
  virtual file system 339
HVC device driver 39
hvc_iucv_allow=, kernel parameter 42
hvc_iucv=, kernel parameter 42
hw_interval
  OProfile attribute 462
  hw_max_interval
  OProfile attribute 462
  hw_min_interval
  OProfile attribute 462
hw_sdbt_blocks
  OProfile attribute 462
hw_trap, qeth attribute 258
hwrng
  cryptographic device node 427
  trng counter 446
hwsampler
  OProfile attribute 461
hwtype
  cryptographic adapter attribute 432
Hyper-Threadin 327
HyperPAV 148
hypervisor
  service level 485
  hypervisor capability 481
hyptps 343
hyptop
  navigate between windows 553
  select data 554
  sort data 554
  units 556
hyptop command
  z/VM fields 555
hyptop, Linux command 551

IBM compatible disk layout 123
IBM Java 365
IBM label partitioning scheme 122
IBM TotalStorage Enterprise Storage Server 121
ica_api.h 439
icainfo, Linux command 489
ICASTats, Linux command 489
IDRC compression 212
IEP 453
if names 4
if_name, qeth attribute 248
IFCC 153
Improved Data Recording Capability compression 212
in_recovery
  zfcp attribute (channel) 172
  zfcp attribute (port) 178, 179
  zfcp attribute (SCSI device) 185
in_recovery, zfcp attribute 169
in-kernel cryptography 449
inbound checksum
  offload operation 252
  inbound checksum, qeth 253
Initial Program Load
  See IPL
initial RAM disk 83
initr
module parameters 29
instruction execution protection 453
Integrated ASCII console applet on HMC 37
interface
MTO 207
network 4
interface control check 153
interface names
ctc 306
mpc 306
qeth 231, 248
storage class memory 201
versus devices 5
vmur 399
interfaces 439
CTC 306
FC-HBA 163
interrupt
cryptographic device attribute 435
invalid_crc_count zfcp attribute 170
invalid_tx_word_count zfcp attribute 170
iocounterbits
zfcp attribute 186
ioctl
protected key device driver 448
iodone_cnt
zfcp attribute (SCSI device) 186
ioerr_cnt
zfcp attribute (SCSI device) 186
iorequest_cnt
zfcp attribute (SCSI device) 186
ip 4
IP address
confirming 250
duplicate 251
takeover 263
virtual 268
IP address takeover, activating and deactivating 264
ip link
command 279
ipa_takeover, qeth attributes 263
IPL 81
displaying current settings 583
IPL devices
for booting 82
preparing 55
IPv6
source VIPA 275
stateless autoconfiguration 231
support for 231
ISO-8859-1 506
isolation, qeth attribute 254
IUCV
accessing terminal devices over 46
authorizations 320
enablement 320
maximum number of connections 320
OPTION MAXCONN 320
iucvconn 34
set up a z/VM guest virtual machine for 43
using on z/VM 39
iucvitty 43
iucvitty, Linux command 43

J
Java, GenWQE 359
Java, GenWQE acceleration 365

K
KB xi
KEK 619
cryptographic API 449
kernel image 83
kernel module 23
aes_s390 450
af_iucv 321
ap 430
appldata_mem 375
appldata_net_sum 375
appldata_os 375
cmm 421
citm 307
dasd_diag_mod 133
dasd_eckd_mod 132
dasd_fba_mod 132
dasd_mod 131
dcssblk 402
des_s390 450
diag288_wdt 116
ghash_s390 450
hmcdrvfs 356
lcs 300
monreader 387
monwriter 381
paes_s390 450
pkey 447
qeth 237
qeth_j2 237
qeth_j3 237
scp_async 479
sha512 450
tape_34xx 208
tape_3590 208
vmlogrdr 392
vmur 399
xpram 216
zfcp 165
kernel panic 101
creating dump automatically after 485
kernel parameter
etr= 350
kernel parameter file
for z/VM reader 25
kernel parameter line
length limit for booting 26, 27
module parameters 29
kernel parameters 23, 83, 350
and zipl 61
channel measurement facility 457
cio_ignores= 652
cmf.format= 457
cmf.maxchannels= 457
cmma= 656
connmode= 41
console= 41
default_hugepagesz= 339
encoding 23
fips= 657
general 651
kernel parameters (continued)
  hmcdrv.cachesize= 355
  hugepages= 339
  hvc_iucv_allow= 42
  hvc_iucv= 42
  maxcpus= 658
  no_console_suspend= 105
  noexec= 453
  nosm= 659
  pci= 19
  possible_cpus= 660
  ramdisk_size= 661
  reboots= 28
  resume= 105
  root= 662
  sctp_con_drop= 42
  sctp_con_pages= 42
  smt= 664
  specifying= 23
  stp= 351
  vdso= 665
  vmhalt= 666
  vmpanic= 667
  vmpoff= 668
  vmreboot= 669
  zipl= 24

kernel source tree
  key encrypting key= 619
  kilo= 619

LCS device driver
  setup= 300
  LDL disk layout= 126
  LGR= 374
  libcard, GenWQE= 359
  libfuse
    package= 640
    libhbaapi-devel= 196
    libica= 426
    libzfcpbaapi= 198
    libzfcpbaapi, package= 198
    libzHW= 359
  lic_version, zfcp attribute= 169
  line edit characters, z/VM console= 54
  line-mode terminal= 43
  control characters= 49
  special characters= 49
  link_failure_count, zfcp attribute= 170

Linux
  as LAN sniffer= 286
  device categories= 7
  Linux commands
    generic options= 489
  Linux device special file
    See device nodes
  Linux guest relocation= 374
  Linux in LPAR mode, booting= 89
  Linux on z/VM
    booting= 84
    reducing memory of= 373

lip_count, zfcp attribute= 170
  list media contents= 357
  listxattr= 357, 640
  LNX1 labeled disk= 126
  load balancing and VIP= 275
  LOADDEV= 86
  LOADNSHR operand= 401
  log file, osasnmpd= 295
  log information
    FCP devices= 174
    logging
      I/O subchannel status= 473
      logical boot device
        base parameters= 63
        helper script= 65
  LOGREC, z/VM record= 391
  long random numbers= 431
  loss_of_signal_count, zfcp attribute= 170
  loss_of_sync_count, zfcp attribute= 170
  lost DASD reservation= 151
  LPAR
    configuration
      storage-class memory= 201
      hardware counters= 466
      I/O devices, configuring= 484
    set up for hardware sampling= 461
  LPAR configuration= 201
  LPAR Linux, booting= 89
  leacdl
    metadata file attribute= 640
    lschp, Linux command= 562
    lscpu, Linux command= 327
    lsmpum, Linux command= 564
    lscss, Linux command= 202, 567
    lsdasd, Linux command= 571
    lshmc, Linux command= 575
mount point (continued)

profs xi
sysfs xi
mt-st, package 212
MTIO interface 207
MTU
qeth 249
multicast_router, value for qeth router attribute 260
multiple subchannel set 11
multithreading 327

N
name
devices
See device names
names
DASD 128
navigate between windows
htop 553
net-snmp 289
network
device drivers 219
network concentrator
examples 282
Network Concentrator 280
network interfaces 4
network names 4
network traffic analyzer
HiperSockets 286
no_auto_port_rescan=
zfcp module parameters 165
no_console_suspend, kernel parameters 105
no_priority_queueing, value for qeth priority_queueing
attribute 244
no_router, value for qeth router attribute 260
node_name
zfcp attribute 169
zfcp attribute (port) 178
node, device
See device nodes
noexec=
kernem parameter 453
non-operational terminals
preventing re-spawns for 44
non-priority commands 51
noresume, kernel parameters 105
nos_count, zfcp attribute 170
nosmt, kernel parameter 659
nowayouts, module parameters 116
NPIV 182
example 174
FCP channel mode 173
for FCP channels 164
removing SCSI devices 192
nr_hugepages
large page support attribute 341
numbers, random 431

O
object ID 289
offline
CHPID 15, 16
devices 9

offload operations
inbound checksum 252
outbound checksum 252
TCP segmentation offload (TSO) 252
OID (object ID) 289
online
CHPID 15, 16
common CCW attribute 9
CPU attribute 329
cryptographic adapter attribute 433
CTCM attribute 310
DASD attribute 140, 141
etr attribute 351
LCS attribute 302
qeth attribute 248
stp attribute 352
tape attribute 209, 210
TTY attribute 48
zfcp attribute 167
opcontrol 461
Open Source Development Network, Inc. 289
openCryptoki, library 439
Operating System Messages applet
emulation of the HMC 51
on HMC 37
operation, tape attribute 211
OProfile
hardware sampling 461
hw_interval attribute 462
hw_max_interval attribute 462
hw_min_interval attribute 462
hw_sdbt_blocks attribute 462
hwsampler attribute 461
initializing 461
starting and stopping 462
OPTION MAXCONN 320
optional properties
DCSS 402
OSA-Express
device driver 223
LAN channel station 299
SNMP subagent support 289
OSA-Express MIB file 290
osasnmvipd
access control 291
checking the log file 295
master agent 289
setup 290
starting the subagent 295
stopping 296
subagent 289
osasnmvipd, OSA-Express SNMP subagent 289
OSDN (Open Source Development Network, Inc.) 289
outbound checksum
offload operation 252
outbound checksum, qeth 253
overlap with guest storage 387

P
PAES 449
paes_s390, kernel module 450
page pool
static 373
timed 374
parallel access volume (PAV) 156
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>poll thread (continued)</td>
<td></td>
</tr>
<tr>
<td>enable using chcrypt</td>
<td>503</td>
</tr>
<tr>
<td>poll_thread</td>
<td></td>
</tr>
<tr>
<td>AP bus</td>
<td>438</td>
</tr>
<tr>
<td>cryptographic adapter attribute</td>
<td>434, 435</td>
</tr>
<tr>
<td>poll_thread=</td>
<td></td>
</tr>
<tr>
<td>module parameter</td>
<td>430</td>
</tr>
<tr>
<td>poll_timeout</td>
<td></td>
</tr>
<tr>
<td>cryptographic adapter attribute</td>
<td>436</td>
</tr>
<tr>
<td>set using chcrypt</td>
<td>503</td>
</tr>
<tr>
<td>port scan</td>
<td></td>
</tr>
<tr>
<td>controlling</td>
<td>175</td>
</tr>
<tr>
<td>port_id</td>
<td></td>
</tr>
<tr>
<td>zfcp attribute (port)</td>
<td>178</td>
</tr>
<tr>
<td>port_id, zfcp attribute</td>
<td>169</td>
</tr>
<tr>
<td>port_name</td>
<td></td>
</tr>
<tr>
<td>zfcp attribute (port)</td>
<td>178</td>
</tr>
<tr>
<td>port_name, zfcp attribute</td>
<td>169</td>
</tr>
<tr>
<td>port_remove, zfcp attribute</td>
<td>180</td>
</tr>
<tr>
<td>port_rescan, zfcp attribute</td>
<td>174</td>
</tr>
<tr>
<td>port_scan,backoff</td>
<td>175</td>
</tr>
<tr>
<td>port_scan_rate_limit</td>
<td>175</td>
</tr>
<tr>
<td>port_state</td>
<td></td>
</tr>
<tr>
<td>zfcp attribute (port)</td>
<td>178</td>
</tr>
<tr>
<td>port_type, NPIV</td>
<td>182</td>
</tr>
<tr>
<td>port_type, zfcp attribute</td>
<td>169</td>
</tr>
<tr>
<td>portno, qeth attribute</td>
<td>246</td>
</tr>
<tr>
<td>ports</td>
<td></td>
</tr>
<tr>
<td>listing</td>
<td>195</td>
</tr>
<tr>
<td>possible_cpus=</td>
<td></td>
</tr>
<tr>
<td>kernel parameter</td>
<td>660</td>
</tr>
<tr>
<td>power attribute</td>
<td></td>
</tr>
<tr>
<td>PCIe</td>
<td>20</td>
</tr>
<tr>
<td>power/state attribute</td>
<td>107</td>
</tr>
<tr>
<td>prandom</td>
<td></td>
</tr>
<tr>
<td>byte_counter attribute</td>
<td>443</td>
</tr>
<tr>
<td>chunksize attribute</td>
<td>443</td>
</tr>
<tr>
<td>errorflag attribute</td>
<td>443</td>
</tr>
<tr>
<td>for non-root users</td>
<td>442</td>
</tr>
<tr>
<td>mode attribute</td>
<td>443</td>
</tr>
<tr>
<td>preferred console</td>
<td>41</td>
</tr>
<tr>
<td>preparing ECKD</td>
<td>135</td>
</tr>
<tr>
<td>preparing FBA</td>
<td>136</td>
</tr>
<tr>
<td>pri=, fstab parameter</td>
<td>106</td>
</tr>
<tr>
<td>prim_seq_protocol_err_count, zfcp attribute</td>
<td>170</td>
</tr>
<tr>
<td>primary_connector, value for qeth router attribute</td>
<td>261</td>
</tr>
<tr>
<td>primary_router, value for qeth router attribute</td>
<td>260</td>
</tr>
<tr>
<td>priority_queueing, value for qeth priority_queueing attribute</td>
<td>245</td>
</tr>
<tr>
<td>priority command</td>
<td>51</td>
</tr>
<tr>
<td>priority_queueing, qeth attribute</td>
<td>244</td>
</tr>
<tr>
<td>prng</td>
<td></td>
</tr>
<tr>
<td>reseed</td>
<td>444</td>
</tr>
<tr>
<td>reseed_limit</td>
<td>444</td>
</tr>
<tr>
<td>processors</td>
<td></td>
</tr>
<tr>
<td>crypto</td>
<td>7</td>
</tr>
<tr>
<td>proofs</td>
<td></td>
</tr>
<tr>
<td>appdata</td>
<td>375</td>
</tr>
<tr>
<td>cio_ignore</td>
<td>653</td>
</tr>
<tr>
<td>magic sysrequest function</td>
<td>50</td>
</tr>
<tr>
<td>VLAN</td>
<td>279</td>
</tr>
<tr>
<td>programming interfaces</td>
<td></td>
</tr>
<tr>
<td>protected key device driver</td>
<td>448</td>
</tr>
<tr>
<td>prot_capabilities</td>
<td></td>
</tr>
<tr>
<td>zfcp attribute</td>
<td>194</td>
</tr>
<tr>
<td>protected key</td>
<td>447</td>
</tr>
<tr>
<td>programming interfaces</td>
<td></td>
</tr>
<tr>
<td>protocol, CTCM attribute</td>
<td>309</td>
</tr>
</tbody>
</table>

Index 703
proxy ARP 267
proxy ARP attributes 240
pseudorandom number
device driver 441
pseudorandom number device driver
ssetup 441
PSW
disabled wait 485
purge, z/VM recording attribute 394
PVMSG 51

Q
QDIO 230
qeth
activating an interface 249
activating and deactivating IP addresses for takeover 264
auto-detection 230
bridge_hostnotify attribute 235
bridge_role attribute 235, 269
bridge_state attribute 235
buffer_count attribute 246
card_type attribute 247
configuration tool 610
deactivating an interface 251
device directories 231
device driver 223
device driver functions 226
displaying device overview 581
enable attribute for IP takeover 264
fake_broadcast attribute 263
group attribute 241
group devices, names of 229
hsuid attribute 269
hw_trap attribute 258
if_name attribute 248
interface names 231
ipa_takeover attributes 263
isolation attribute 254
layer 2 229
layer 3 229
layer2 attribute 231, 243
MTU 249
online attribute 248
portno attribute 246
priority_queueing attribute 244
problem determination attribute 240
proxy ARP attributes 240
recover attribute 252
removing modules 238
route4 attribute 260
route6 attribute 260
sniffer attributes 240
subchannels 230
summary of attributes 239, 240
switching the discipline 238
TCP segmentation offload 262
ungroup attribute 242
VIPA attributes 240
qeth interfaces, mapping 5
QETH performance statistics 257
qetharp, Linux command 608
qethconf, Linux command 610
qethiqoat, Linux command 613
query HPF
DASD 154
queue_depth=, zfcp module parameters 165
queue_ramp_up_period, zfcp attribute 188
queue_type
zfcp attribute (SCSI device) 186
queueing, priority 244
R
RAM disk, initial 83
ramdisk_size=, kernel parameter 661
random number
device driver 441, 445
random numbers
reading 442, 446
raw_track_access, DASD attribute 149
raw-track access mode 640
RDMA 19
read monitor data 372
readelf, Linux command 453
readlink, Linux command 5
readonly
DASD attribute 156
reboot
kernel parameters 28
recovery
metadata file attribute 640
record layout
z/VM 391
recording, z/VM recording attribute 393
recover
PCIe attribute 21
recover, lcs attribute 304
recover, qeth attribute 252
recovery, CTC interfaces 313
Red Hat Enterprise Linux
documentation website 687
reflective relay mode 254
relative port number
qeth 246
Remote Direct Memory Access (RDMA) 19
Remote Spooling Communications Subsystem 631
Removable media, loading Linux 95
remove channel path
DASD 153
remove, DCSS attribute 409
removing
 cryptographic modules 438
request processing
 cryptographic 427
request_count
 cryptographic adapter attribute 432
rescan
zfcp attribute (SCSI device) 190
reseed
random attribute 443
prng 444
reseed_limit
random attribute 443
prng 444
reseed_limit=, module parameters 441
reservation state
DASD 152
reservation_policy, DASD attribute 151
reset_statistics
zfcp attribute 170
respawn prevention 44
resume 103

704 Device Drivers, Features, and Commands - Red Hat Enterprise Linux 7.5
resume=, kernel parameters 105
reuse 216
rev
zfcp attribute (SCSI device) 186

RFC
  1950 (zlib) 359
  1951 (deflate) 359
  1952 (gzip) 359
Rivest-Shamir-Adleman 425
Ro, kernel parameter 662
RoCE 19
roles
  zfcp attribute (port) 178
root=, kernel parameter 663
route4, qeth attribute 260
route6, qeth attribute 260
router
  IPv4 router settings 260
  IPv6 router settings 260
RPM
genwqe-tools 362
genwqe-zlib 362
libfuse 640
libhbaapi-devel 196
libhugetlbfs 339
libica 426
libzfcphbaapi 198
mt-st 212
openCryptoki 439
oprofile 461
s390utils 489
gs3_utils 587
src_vipa 275
util-linux 327
RSA 425
RSA exponentiation 425
RSCS 631
rx_frames, zfcp attribute 170
rx_words, zfcp attribute 170

S
s_id, zfcp attribute 173
S/390 hypervisor file system 343
defining access rights 346
directory structure 343
  LPAR directory structure 343
  updating hypfs information 347
z/VM directory structure 344
s390dbf 374
ts390utils, package 489
sample_count, cmf attribute 459
sampling facility
  reading data 467
save, DCSS attribute 408
SCLP_ASYNC 479
SCLP_ASYNC device driver 479
scil_con_drop=, kernel parameter 42
scil_con_pages=, kernel parameter 42
SCM 205
scm_block=, module parameters 202
script
  base device 65
SCSI
data consistency checking 194
device nodes 161
  multipath devices 163
SCSI device
  automatically attached, configuring 182
  configuring manually 182
SCSI devices
  in sysfs 184
  information in sysfs 185
  partitioning 163
  removing 192
  sysfs structure 160
SCSI devices, in sysfs 184
SCSI system dumper 68
SCSI tape
lstape data 589
scsi_host_no, zfcp attribute 184
scsi_id, zfcp attribute 184
scsi_level
  zfcp attribute (SCSI device) 186
scsi_logging_level, Linux command 616
scsi lun, zfcp attribute 184
scsi_target_id
  zfcp attribute (port) 178
SCSI-over-Fibre Channel
  See zfcp
SCSI-over-Fibre Channel device driver 159
SCSI, booting from 86, 90, 92
SE (Support Element) 82
secondary unicast 227, 228
secondary_connector, value for qeth router attribute 261
secondary_router, value for qeth router attribute 261
seconds_since_last_reset
  zfcp attribute 170
seglist, DCSS attribute 405
segmentation offload, TCP 262
send files
  vmur command 638
send files to z/VSE
  vmur command 639
senders=, module parameter 415
serial_number, zfcp attribute 169
service levels
  reporting to IBM Support 485
service utility
cpuplugd 510
set, CPI attribute 477
setup
  LCS device driver 300
  source VIPA 275
  standard VIPA 272
setbattr 507
sg_inq, Linux command 587
sg3_utils, package 587
SHA-1 449
SHA-256 449
SHA-512 449
SHA-512
  in-kernel cryptography 449
  sha512, kernel module 450
shared, DCSS attribute 407
shutdown actions 109
SIE capability 481
Simple Network Management Protocol 289
simultaneous multithreading 327
sizes=, module parameter 216
SMGS_ID 417
SMGS_SENDER 417
smguiucv_app
  device driver 415
SMT 327

Index 705
subchannels (continued)
    in sysfs 12
    LCS 299
    qeth 230
support
    AF_IUCV address family 319
Support Element 82
supported_classes
    zfcp attribute (port) 178
    supported_classes, zfcp attribute 169
    supported_speeds, zfcp attribute 169
    suspend 103
swap partition
    for suspend resume 105
    priority 106
swapping
    avoiding 373
SYMPTOM, z/VM record 391
syntax diagrams 677
syntax overview
    zipl 56
sysfs 7
    channel subsystem view 12
device view 12
device view by category 11
device view by drivers 11
    FCP devices 160
    information about SCSI devices 185
    representations of SCSI devices 184
    SCSI devices 160
sysfs attribute
    cm_enable 14
    state 335
sysplex_name, CPI attribute 475
system states
    displaying current settings 586
system time protocol 349
    system_time_protocol 349
system_level, CPI attribute 476
system_name, CPI attribute 475
    system_type, CPI attribute 476
    systemd 44

T

T10 DIF 195
tape
    blocksize attribute 211
    booting from 84, 90, 94
cmb_enable attribute 210
cutype attribute 211
device names 206
device nodes 207
devtype attribute 211
display support 623
displaying overview 587
encryption support 619
IDRC compression 212
loading and unloading 213
medium_state attribute 211
MTIO interface 207
    online attribute 209, 210
operation attribute 211
    state attribute 211
    uid attribute 21
tape device driver 205
tape devices
  typical tasks 209
tape390_crypt, Linux command 619
tape390_display, Linux command 623	
tar command, acceleration 364
TCP segmentation offload 262
TCP segmentation offload (ISO)
  offload operation 252
TCP/IP
  ARP 235
  DHCP 285
  point-to-point 305
  service machine 306
TDEA 441
TDES 441
  in-kernel cryptography 449
TERM, environment variable 43
  terminal
  3270, switching the views of 47
  accessing over IUCV 46
  CCW, switching device on- or offline 48
  line-mode 43
  mainframe versus Linux 35
  non-operational, preventing re-spawns for 44
  provided by the 3270 terminal device driver 44
  terminals
  escape character 53
tgid_bind_type, zfcp attribute 169
thread_siblings
  CPU sysfs attribute 330
time
  command 364
  cpuplugd keyword
  use with historical data 517
time-of-day clock 349
time, command 364
timed page pool 374
  reading the size of the 422
  timed page pool size
  setting to avoid guest swapping 484
  timeout
  zfcp attribute (SCSI device) 191
timeout for LCS LAN commands 302
TOD clock 349
Triple Data Encryption Standard 441
tiple DES 441
tng
  byte_counter attribute 446
TRNG device driver
  setup 445
troubleshooting 483
  true random numbers
  reading 446
true random-number device driver
  setup 445
TSO
  offload operation 252
TTY
  online attribute 48
ttyr
  run 44
tunedasd, Linux command 625
tuning automatic port scans 175
tx_frames, zfcp attribute 170
tx_words, zfcp attribute 170
type
  cryptographic adapter attribute 432
  zfcp attribute (SCSI device) 186
  type, CTCM attribute 309
U
udev
  DASD device nodes 128
  handling CP special messages 417
uevent 417
uid
  DASD attribute 157
  PCIe attribute 21
ungroup
  CTCM attribute 308
  LCS attribute 301
  qeth attribute 242
unit_add, zfcp attribute 182
unit_remove, zfcp attribute 193
unloading
  cryptographic modules 438
  updating information
    S/390 hypervisor file system 347
    USB storage, HMC 355
    USB-attached storage, loading Linux 95
use_diag
  DASD attribute 157
  use_diag, DASD attribute 138
user terminal login 44
user:dsor
  extended attribute for z/OS data set 640
user:recl
  extended attribute for z/OS data set 640
user:recfm
  extended attribute for z/OS data set 640
using SCM devices with 203
V
VACM (View-Based Access Control Mechanism) 291
vdso=, kernel parameter 665
vendor
  DASD attribute 157
  zfcp attribute (SCSI device) 186
VEPA mode 254
vfn
  PCIe attribute 21
view
  channel subsystem 12
  device 12
  device by category 11
  device by drivers 11
View-Based Access Control Mechanism (VACM) 291
VINPUT 51
CP command 53
VIPA (virtual IP address)
  attributes 240
  description 268, 272
  example 273
  high-performance environments 275
  source 275
  static routing 272
  usage 272
VIPA, source
  setup 275
VIPA, standard
  adapter outage 273
  setup 272
zfcp (continued)
in_recovery attribute 169
in_recovery attribute (channel) 172
in_recovery attribute (port) 178, 179
in_recovery attribute (SCSI device) 185
invalid_crc_count attribute 170
invalid_tx_word_count attribute 170
iocounterbits attribute 186
iodone_cnt attribute (SCSI device) 186
iorequest_cnt attribute (SCSI device) 186
lic_version attribute 169
link_failure_count attribute 170
lip_count attribute 170
loss_of_signal_count attribute 170
loss_of_sync_count attribute 170
maxframe_size attribute 169
model attribute (SCSI device) 186
node_name attribute 169
node_name attribute (port) 178
nos_count attribute 170
online attribute 167
peer_d_id attribute 169
peer_wwwn attribute 169
peer_wwpn attribute 169
permanent_port_name attribute 169, 173
physical_s_id attribute 173
port_id attribute 169
port_id attribute (port) 178
port_name attribute 169
port_name attribute (port) 178
port_remove attribute 180
port_rescan attribute 174
port_state attribute (port) 178
port_type attribute 169
prim_seq_protocol_err_count attribute 170
protocol_capabilities attribute 194
queue_max_depth attribute 188
queue_ramp_up_period attribute 188
queue_type_attribute (SCSI device) 186
rescan attribute (SCSI device) 190
reset_statistics attribute 170
rev_attribute (SCSI device) 186
roles_attribute (port) 178
rx_frames_attribute 170
rx_words_attribute 170
s_id_attribute 173
scsi_host_no_attribute 184
scsi_id_attribute 184
scsi_level_attribute (SCSI device) 186
scsi_lun_attribute 184
scsi_target_id_attribute (port) 178
seconds_since_last_reset_attribute 170
serial_number_attribute 169
speed_attribute 169
state_attribute (SCSI device) 191
supported_classes_attribute 169
supported_classes_attribute (port) 178
supported_speeds_attribute 169
tgid_bind_type_attribute 169
timeout_attribute (SCSI device) 191
tx_frames_attribute 170
tx_words_attribute 170
type_attribute (SCSI device) 186
unit_add_attribute 182
unit_remove_attribute 193
vendor_attribute (SCSI device) 186
zfcp (continued)
wwpn_attribute 173, 184
wwpn_attribute (SCSI device) 186
zfcp_access_denied_attribute (SCSI device) 186
zfcp_failed_attribute (SCSI device) 189
zfcp_in_recovery_attribute (SCSI device) 186, 189
zfcp_HBA_API 163
zfcp_HBA_API_library 198
zfcp_module_parameters
datarouter= 165
dbsize= 165
dif= 165
no_auto_port_rescan= 165
queue_depth= 165
zfcp_traces 165
zfcp_access_denied
zfcp_attribute (SCSI device) 186
zfcp_diskConfigure 193
zfcp_failed
zfcp_attribute (SCSI device) 189
zfcp_in_recovery
zfcp_attribute (SCSI device) 186, 189
zfcp_ping 198
zfcp_show 198
zipl
and kernel parameters 61
base functions 55
base parameters 63
boot device 59
configuration file 76
Linux command 55
menu configurations 77
modes 56
parameters 72
syntax overview 56
zipl boot menu 35
ZIPLCONF, environment variable 76
zkey, Linux command 645
ZLIB_CARD, environment variable 362
ZLIB_DEFLATE_IMPL, environment variable 362
ZLIB_INFLATE_IMPL, environment variable 362
ZLIB_TRACE, environment variable 362
zlib, GenWQE 359
zlib, RFC 1950 359
znetconf, Linux command 648
Readers’ Comments — We'd Like to Hear from You

Linux on Z and LinuxONE
Device Drivers, Features, and Commands
on Red Hat Enterprise Linux 7.5

Publication No. SC34-2710-05

We appreciate your comments about this publication. Please comment on specific errors or omissions, accuracy, organization, subject matter, or completeness of this book. The comments you send should pertain to only the information in this manual or product and the way in which the information is presented.

For technical questions and information about products and prices, please contact your IBM branch office, your IBM business partner, or your authorized remarketer.

When you send comments to IBM, you grant IBM a nonexclusive right to use or distribute your comments in any way it believes appropriate without incurring any obligation to you. IBM or any other organizations will only use the personal information that you supply to contact you about the issues that you state on this form.

Comments:

Thank you for your support.
Submit your comments using one of these channels:
• Send your comments to the address on the reverse side of this form.
• Send your comments via email to: eservdoc@de.ibm.com

If you would like a response from IBM, please fill in the following information:

Name

Address

Company or Organization

Phone No.

Email address