Linux on System z

Oracle Cluster File System Version 1.2, shared file system for Linux on IBM System z

Linux on System z

Oracle Cluster File System Version 1.2, shared file system for Linux on IBM System z

Before using this information and the product it supports, read the information in “Notices” on page 21.

Edition notices

Contents

About this publication

Authors

Margaret Phillips

Dr. Juergen Doelle

How to send your comments

Your feedback is important in helping to provide the most accurate and highest quality information. If you have any comments about this publication, send your comments using IBM® Resource Link™ at http://www.ibm.com/servers/resourcelink. Click **Feedback** on the navigation pane. Be sure to include the name of the publication, and the specific location of the text you are commenting on (for example, a page number or table number).

Chapter 1. Introduction

This paper studies the Oracle Cluster File System Release 2 (OCFS2) for Linux® on IBM System z®, based on an implementation of OCFS2 in a test lab environment, for a clustered database.

The name 'OCFS2 study' is used in place of the full name 'OCFS2 Version 1.2 shared file system for Linux on IBM System z' for the remainder of this document.[1]

Novell SUSE Linux Enterprise Server (SLES) 10 SP1 on an IBM System z10™ is used. The intent is to provide insight for an enterprise customer who is considering using the OCFS2 shared file system in an IBM System z environment. The following IBM System z features are discussed, in terms of how they relate to OCFS2:

- Servers in LPARs
- HiperSockets™ connections
- Disk storage on large external storage units

This paper is organized into these sections:

1. This paper is intended to provide information regarding OCFS2. It discusses findings based on configurations that were created and tested under laboratory conditions. These findings may not be realized in all customer environments, and implementation in such environments may require additional steps, configurations, and performance analysis. The information herein is provided "AS IS" with no warranties, express or implied. This information does not constitute a specification or form part of the warranty for any IBM products.

Chapter 2. About OCFS2

OCFS2 is described in detail, including its history, present release levels, compatibility issues, and hardware and software requirements.

Overview of OCFS2

OCFS2 is a clustered file system for use with Linux. OCFS2 allows multiple users on physically separated servers to read and write to the same files on the same physical disks at the same time. Access to shared files is coordinated, so that at any moment only one server is actually writing to a block of the file. The integrity of the write operation, the file contents, and the file system structure are protected.

The challenge to shared file systems in general is to make it possible for multiple servers to simultaneously mount the same file structure and to open the same files with read and write access, a scenario that in an unshared file system would lead to corruption of the file and the file system structure. OCFS2 uses a distributed lock manager to organize access and a network interconnect to share the information between clustered servers when a file is changed. When a file is resident in the memory of one of the servers in the cluster, its view of the file content does not change, and that version can be written to the shared file.

A shared file system on clustered servers can be the foundation of a clustered database design, but clustered database instances also communicate with each other, sharing cached data and metadata. Those types of cluster functions are typically part of a database application's cluster stack, and not that of OCFS2. The operation of an OCFS2 file system is managed using several services in the O2CB stack. The O2CB stack must be loaded, started, and brought online before anything can be done with the OCFS2 clustered file system.

Shared file systems are also used for creating high availability (HA) systems, where redundant access to data or executable files in storage provides failover access to SAN storage or local disk storage.

The shared file system can be used simply for the economy that comes from maintaining data in a single location rather than multiple locations.

This study is based on OCFS2 Version 1.2 distributed by Novell, with the SUSE Linux Enterprise Server (SLES) kernel version 10 SP1 on IBM System z. This version of OCFS2 was developed to support clustered databases in an enterprise environment.

The Novell distributions of OCFS2 are fully supported by Novell. OCFS2 is also available on Red Hat Enterprise Linux and Oracle Enterprise Linux, and both of those distributions are supported by Oracle when used with Oracle products.

History of OCFS2

It is helpful to understand the history of OCFS2 and the changes that it has undergone.

Oracle Cluster File System (OCFS), the predecessor of OCFS2, was introduced in 2003. OCFS2 was created by Oracle as the file system to be used by Oracle Real

Application Clusters (RAC). RAC is a clustered database designed to have multiple instances on separate servers or nodes, with shared simultaneous access to data in real time.

OCFS2 Version 1.2 was released in 2006 with the Enterprise Linux distributions including SLES 9 SP4, SLES 10 SP1, RHEL 4 Update 4, and RHEL 5. For Novell's SUSE Linux Enterprise Server, OCFS2 Version 1.2 is bundled with SLES 9 SP4 and SLES 10 SP1, whereas OCFS2 Version 1.4 is bundled with SLES 10 SP2. At the time of this writing, Red Hat packages for OCFS2 Version 1.2 were available for RHEL 4 and RHEL 5 kernels, which are distributed by Oracle. At the time of this writing, packages for OCFS2 Version 1.4 were available for RHEL 5 Update 2. The OCFS2 file system itself is distributed as a part of the kernel. The tools needed to manage OCFS2 (mkfs.ocfs2, fsck.ocfs2, ocfs2console, and others) are in separate packages named ocfs2-tools and ocfs2console. Here the contents of the various distributions differs. For SLES 11, these packages are part of the SUSE Linux High Availability Extension package.

OCFS2 Version 1.2 can be used by Oracle RAC for:

- Data storage
- Files for cluster management referred to as Voting disks
- The Oracle Cluster Registry (OCR)
- Binary code for Oracle products

One use of shared disk storage is to store data for clustered databases, which allow multiple users on dispersed database servers to access a common database.

There are two different approaches to implement shared disk storage on clustered applications such as a cluster database:

1. Implementation of shared storage in the file system, such as OCFS2
2. Implementation of shared storage in the application level, for example Oracle's Automatic Storage Manager (ASM).

Comparing these two approaches for any given project requires considering the whole scope of the end use, since one is a Linux file system and the other is an application that has many functions. With ASM, the disks are owned by the application and files are served while many disk operations (including striping and coordinating shared access) occur within the application.

The latest OCFS2 version is OCFS2 Version 1.4. Files from OCFS2 Version 1.2 can be mounted with OCFS2 Version 1.4 nodes, but the newer features for high availability are not turned on by default. One significant restriction is that a volume cannot be mounted on OCFS2 Version 1.2 on one node and mounted on OCFS2 Version 1.4 on another node at the same time.

OCFS2 Version 1.4 is optimized for a high availability (HA) package from Novell named SUSE Linux High Availability Extension, which uses the potential for redundant access of the OCFS2 file system to reduce downtime and maintain business continuity.

Versatility of OCFS2

Businesses of all sizes are looking for ways to be productive with the tools already at hand. OCFS2 Version 1.2 is included with SLES 9 and SLES 10, making it an economical resource for users who already have either of these Linux distributions.

See the following Oracle development Web site, which contains design documents and test information:

http://oss.oracle.com/osswiki/OCFS2

Systems and connection methods for OCFS2

The servers and their communication methods for the OCFS2 study are described in detail.

The cluster of two Linux servers created with SLES 10 SP1 is installed on separate Logical Partitions (LPARs) of an IBM System z10. A Linux server installed on an LPAR can be configured with dedicated or shared resources, including processor, memory, and I/O devices. Each Linux LPAR uses two dedicated processors and 2 GB of memory.

Each Linux server uses dedicated storage for their SLES 10 SP1 operating system and user home directories on non-shared disks formatted with EXT3, and shared disks formatted with OCFS2. All of the physical disks were RAID arrays on an IBM System Storage™ DS8000®, large external disk storage unit.

The communication channel between the servers for sharing the disks is used to communicate cluster membership status and other cluster-related messages between cluster members. Communication channels can be implemented with a wired Ethernet network (for example, a LAN), or on the IBM System z with a HiperSockets connection. IBM System z HiperSockets are very fast internal channels between LPARs on the same physical machine, which operate directly in the memory from the IBM System z10 Central Electronic Complex (CEC), without any external cables or switches. The short latency time of HiperSockets communications compared with other internode communications makes IBM System z suitable for clustered applications.

Storage server and disk configuration for OCFS2

This OCFS2 study used DASD type disk storage, which uses the IBM ECKD™ format and FICON® channels.

Using Fibre Channel Protocol (FCP) connected SCSI disks would have been also appropriate for this system. The DASD devices were attached with four FICON Express4 channels (4 Gb/sec link speed) using switched connections to four host adapters on the IBM System Storage DS8000 storage unit.

When selecting the disks for shared storage on the IBM System Storage DS8000, they were spread out as much as possible among ranks (areas) of the disk array drawers, to prevent disk access from being concentrated in any one area. The IBM System Storage DS8000 consists of two server complexes, each having two power processors and separate caches each with half of the total cache size. The disk storage was selected from the ranks pre-configured on this IBM System Storage DS8000, to use the caches from both processor complexes.

More information about performance considerations for databases for Linux on IBM System z is available at the following Web site:

http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/perf/Performance_considerations_for_databases_on_Linux_on_System_z.pdf

See the documentation for your storage control unit to determine the disk layout that gives the best performance. For the IBM System Storage DS8000 series, refer to IBM System Storage DS8000 Performance Monitoring and Tuning, an IBM Redbooks® publication, available at:

http://www.redbooks.ibm.com/abstracts/sg247146.html

If I/O latency is important to the undertaking, and striped storage is to be used, implement striping at the level of the physical storage server, if it is available. The striping, therefore, must be configured for the disk devices that are going to be used for shared storage before they are introduced to Linux as devices. The IBM System Storage DS8000 storage server has the facility to stripe disk storage, creating logical disks that are physically spread across ranks of physical disk arrays in the unit. Striping was introduced with IBM System Storage DS8000 microcode level r3 code 63.0106.0, and can be upgraded to the appropriate level.

On an enterprise system, being able to expand volumes dynamically is an important convenience supplied by the Logical Volume Manager (LVM), so administrators must give extra thought to prepare for expanding a system without it. The problem appears to be that the LVM must be developed with the same cluster awareness of which nodes are alive in a cluster as O2CB has for managing the file system. There is not an existing version of clustered LVM that works with OCFS2. OCFS2 Version 1.4 was announced as having moved towards being able to resize volumes dynamically, but at the time of this writing, it is not fully supported.

The O2CB heartbeat and services stack

What makes OCFS2 unique is its stack of five file system services, named O2CB.

O2CB manages the shared file access within a cluster of servers, and although these services can be transparent to a user after setup, the O2CB stack must be loaded and brought online at system restart time. An OCFS2 file looks like any other file.

O2CB has these five services:

Node manager (NM)
Keeps track of all nodes in the cluster.

Heart beat (HB)
Issues node up and node down notification when nodes join or leave the cluster.

TCP Provides communication between the nodes.

Distributed Lock Manager (DLM) and Distributed Lock Manager File System (DLMFS)
These two services ensure the consistency of the clustered file system:

DLM Keeps track of all locks, lock holders, and lock status.

DLMFS
Provides the user space interface to DLM in the Linux kernel.

CONFIGFS
Performs filesystem-based management of kernel objects, known as config_items. The mount point is directory `/config`.

The O2CB heartbeat is created by cluster members writing to a shared control area at regular intervals, using the designated interconnect, as proof that the node is alive. The default heartbeat rate is one heartbeat every two seconds. Even in a clustered file system with only one node, the heartbeat is maintained.

The Node Manager coordinates the heartbeat processes and responds when nodes join, drop off, lose communication, or change status, whether intentional or not, within the cluster. The Node Manager can trigger recovery. A key function of the Node Manager is to make it possible to free resources in the event of a node dropping off.

The Distributed Lock System is the key to OCFS2's coordination of shared access to storage by multiple nodes. The DLS is a hierarchical locking system that allows enqueueing for resources from high level to granular. Locking is administered by the DLM.

Chapter 3. Installing and customizing OCFS2

You must install OCFS2 and then set up the cluster and disks for file system access.

Installing OCFS2 packages

The two OCFS2 Version 1.2 packages that are installed with Novell SUSE Linux Enterprise Server (SLES) 10 SP1 are named ocfs2-tools and the ocfs2console.

It is possible to do every installation and configuration step using Linux commands. However, the ocfs2console, a GUI program, is advantageous for performing Step 1 in "Creating the OCFS2 cluster environment." This step is to create and propagate identical OCFS2 configurations to all the nodes in the cluster. The ocfs2console ensures that the cluster configuration files work together. The installer code level must match the code level of the file system.

The ocfs2console can be used for these tasks:

1. Build the cluster of nodes.
2. Format the shared devices.
3. Mount and unmount OCFS2 volumes.
4. Add or remove nodes to the cluster.
5. Label disks.
6. Add slots to disk clusters so that more nodes can access the disk.

Creating the OCFS2 cluster environment

Preparing to implement the OCFS2 file system and bring it online with a heartbeat requires creating a set of servers (nodes), preparing a set of empty disks with Linux, and making OCFS2 available on each server.

The ocfs2console creates one configuration file for the first node in the cluster, and propagates it to all the nodes using ssh. The propagation requires that the root user on the nodes can use ssh on each other node without entering a password or a passphrase.

To create a cluster for use by OCFS2, complete these steps:

1. As the root user, generate a pair of public and private keys using the `keygen` command:
   ```
   ssh-keygen -t rsa
   ```
2. Accept the default locations and respond to the request for a passphrase with an appropriate password. A pair of authentication keys in the `~/.ssh/id_rsa` and `~/.ssh/id_rsa.pub` directories will be created.
3. Append the contents of the `~/.ssh/id_rsa.pub` public key file to the `~/.ssh/authorized_keys` file, and then append the same public key file to all the nodes in the OCFS2 cluster. It might be necessary to create the `~/.ssh/authorized_keys` file.
4. Before proceeding, issue the `ssh` command to use each of these access keys, because the first time that they are used the system asks for an interactive response 'y' from the operator to add the system to the list of known hosts.

5. As the root user:
   a. Open a graphical interface such as VNC.
   b. Open the ocfs2console.
   c. Create the OCFS2 cluster on the installer by selecting **Cluster → Configure Nodes**. This builds the `/etc/ocfs2/cluster.conf` file. Either accept the default name for the cluster, which is OCFS2, or choose a name.
6. From the Task menu, select **Add**, and for each node type the name of the node, IP address to be used by OCFS2, and a port number of 7777, which is considered a default port for cluster communications.

   The node name must match the host name, but the extended domain name is not needed. As an example, use **myserver** instead of **myserver.xyz.com**. The IP address does not need to be the one associated with that host name. Any valid IP address on that node can be used. OCFS2 does not match the node name (host name) with the specified IP address.
7. When all the nodes are added, select **Propagate** and click **OK**.

   The creation of the initial cluster of servers that share the OCFS2 file system is now complete.

Cluster members can be added or deleted at a later date in the same manner that the cluster is created. The cluster is now ready to be brought online in preparation for adding the disks for shared storage.

Controlling O2CB with the `/etc/init.d/o2cb` utility

The utility `/etc/init.d/o2cb` is used to manage the cluster services, also known as the O2CB stack.

To bring the cluster online, run the following command, where *MyCluster* is the name of the cluster:

```
/etc/init.d/o2cb online MyCluster
```

To stop O2CB, run the following command:

```
/etc/init.d/o2cb stop
```

To show the status of O2CB, run the following command:

```
/etc/init.d/o2cb status
```

The output of the status command is similar to this output:

```
Module "configfs": Loaded
Filesystem "configfs": Mounted
Module "ocfs2_nodemanager": Loaded
Module "ocfs2_dlm": Loaded
Module "ocfs2_dlmfs": Loaded
Filesystem "ocfs2_dlmfs": Mounted
Checking O2CB clusterChecking O2CB cluster mycluster: Online MyCluster: Online
Checking O2CB heartbeat: Active
```

Note that the cluster in the example is online, which would not be possible if the other conditions were not either `Loaded` or `Mounted`.

Preparing disks for use with OCFS2

Prepare the disks for use on the cluster, if this task has not already been done.

Linux on IBM System z is not aware of disk devices until they are brought online, which is when they are assigned a Linux file name. Linux file names are `/dev/dasd`*x*, or `/dev/scd`*x* for a SCSI device, where *x* identifies the disk in Linux.

To prepare the disks for use with OCFS2 on SUSE Linux Enterprise Server, complete these steps:

1. If it is necessary to bring the disks you select for the cluster online, issue the `chccwdev` command as shown, where 722c is the disk device number from the disk storage unit:

   ```
   chccwdev -e 722c
   ```

2. Verify the results of the `chccwdev` command to see which disks are online, by issuing this command:

   ```
   lscss |grep yes
   ```

3. Issue the `lsdasd` or `lscss` command to correlate the Linux names with the device numbers.
4. If the disk has not been prepared with low-level formatting for DASD, issue the `dasdfmt` command, where *x* identifies the disk:

   ```
   dasdfmt -f /dev/dasdx -b 4096 -p
   ```

5. Partition the disks, by issuing this command:

   ```
   fdasd -a /dev/dasdx
   ```

6. Check to see that the disks are partitioned by viewing the file `/proc/partitions`, with this command:

   ```
   cat /proc/partitions
   ```

7. Verify that none of the devices are mounted at this point in the installation, and that the O2CB cluster is online. The ocfs2console should display the disks as online and visible to the server in the cluster that is running the ocfs2console.
8. Format the disks for OCFS2 on the ocfs2console by clicking **Tasks → Format**. Select the disks one at a time from a drop-down list of available devices on the ocfs2console.
9. Specify these values for each disk device:

   Cluster size
   Unit of space, in bytes, that is allocated for a file. Select a value that is larger than the blocksize. Supported values are 4 KB through 1 MB. This study chose the largest unit, 1 MB, for data storage.

   Blocksize
   Smallest unit of space, in bytes, addressable by the file system (managed with the metadata). Supported sizes are 4 KB, 8 KB, 16 KB, 32 KB, 64 KB, 256 KB, and 1 MB. This study chose 4 KB, because it is the Linux block size.

   Number of slots
   Number of nodes that can use the volume concurrently. Values range from 1 to 255. The number of slots can be increased later, but not decreased, by using the task bar on the ocfs2console. This study chose two slots for a two-node system.

Format the disks only on one node, which can be done with a Linux command as shown. This example formats the disk named `/dev/dasdg1` with a blocksize of 4 KB and cluster size of 1 MB, and creates two slots for use by a two-node cluster:

```
mkfs.ocfs2 -b 4k -C 1M -N 2 -L MyLabel /dev/dasdg1
```

As shown above, the disks accept labels, and the ocfs2console has a drop down task menu to put a label on each disk. This optional feature might help the administrator to keep track of the disks being mounted on multiple nodes. However, if labels are not used for all tasks such as backing up or in the `zipl.conf` file (described in "Too many DASDs accessible to Linux for read" on page 16), then using labels in some situations will not simplify mapping disks; it creates another name field to map.

See "Managing shared disks" on page 15 for strategies to keep track of shared disks on a storage control unit with IBM System z.

Starting OCFS2 at system restart time

It is convenient to have OCFS2 loaded and start automatically at system restart time.

To set up an automatic start and stop of OCFS2, complete the following steps:

1. Modify the `configure.conf` file on each node in the cluster using O2CB, by running the following command:
   ```
   /etc/init.d/o2cb configure
   ```
2. Respond 'y' to the following question:
   ```
   Load O2CB driver on boot (y/n) [n] : y
   ```
3. To make sure that OCFS2 is online, run the following command:
   ```
   chkconfig -l ocfs2
   ```
4. Verify that Linux run levels 2, 3, 4 and 5 are all set 'on'.
   ```
   ocfs2     0:off   1:off   2:on    3:on    4:on    5:on    6:off
   ```

Mounting shared devices and using the `/etc/fstab` file

Mounting shared devices can be done from the ocfs2console, or from the command line.

The shared disks must be mounted on each node individually, but it is not necessary for all the disks to be mounted on all the nodes. If a particular node does not need access to a particular disk, there is no need to have that disk mounted on that node. When disks are mounted on multiple nodes, you can verify the shared file behavior by creating a file with an editor and saving it, which makes it immediately visible from any of the nodes.

The following example, from file `/etc/fstab`, shows the required mount options for OCFS2 disks that will be mounted at system restart time. The first device, `/dev/dasdf1` is a partitioned data disk. The second device, `/dev/dasdd1`, is for disks that are shared sources of binary executables.

Issue this command:

```
cat /etc/fstab
```

The output is similar to these lines:

```
/dev/dasdf1 /datadisk1 ocfs2   _netdev,datavolume,nointr  0     0
/dev/dasdd1 /bin_disk  ocfs2   _netdev                    0     0
```

The mounting options from the `cat /etc/fstab` command output are defined in the following list:

_netdev
Ensures that the OCFS2 volume is not mounted before the networking structure is up, and ensures that there is an unmount before shutting down the network.

datavolume
Applies only to data volumes, and every type of file usage except shared binaries. On a clustered database such as Oracle Real Application Clusters (RAC), the datavolume includes the Cluster Registry and Voting Disks. The datavolume allows direct I/O access to the files.

nointr Prohibits interrupts, and is applied to the same type of data files as the datavolume option.

When Oracle RAC Voting Disks, and OCR (Oracle Cluster Registry) disks are installed on OCFS2, the disks require the same mounting options as datavolumes: `_netdev`, `datavolume`, `nointr`.

Chapter 4. Maintaining OCFS2

After OCFS2 is installed and configured, the cluster is defined, the disks are allocated, and everything has been started, you might need to make changes to ensure that this setup remains current and meets your needs.

After OCFS2 is installed, it requires very little maintenance. However, it is good to know that the Linux file checker, named `fsck.ocfs2`, can be used to check and correct problems in an OCFS2 target, typically a disk device. By default, the `fsck.ocfs2` utility uses the cluster services. To verify that the specified file or device is not mounted on any node in the cluster, issue the following command:

```
fsck.ocfs2  device
```

The `fsck.ocfs2` utility searches for file system errors and tries to correct them if the invoker of the command selects an option to do so. The complete syntax and explanation of return codes is on the Linux man page for the `fsck.ocfs2` utility.

The ocfs2console has file checking and repair as tasks on the drop-down task menu, and checks the disk that the user selects.

Managing shared disks

Implementing OCFS2 for a project requires developing a strategy to keep track of the disks, that are most likely known by different names and addresses when referenced from different servers, or by physical addresses on the storage units.

Backup and recovery procedures can require cross-referencing of shared DASD, and clustering applications can require persistent names that are always the same after system restart, and are also the same on all nodes. Incorrect identification of shared disks can cause them to be overwritten or lost.

Linux assigns names to all devices it discovers during system restart, in the sequence in which Linux discovers them, starting with `dasda` or `sda`, and continuing to assign names in alphabetical order Even with few disks used from a SAN, the order can change from one system restart to the next. For example, if one disk in the sequence becomes unavailable, then all the disks that follow shift to a different name in the series. The naming order might change in a way that affects the individual nodes differently.

One method to be sure which disk is being mounted with file `/etc/fstab` is to use the `/dev/disk/by-id` name for a partitioned disk, which looks like this:

```
ccw-IBM.750000000L5611.2bea.b4-part1
```

The Linux device name after creating a single partition looks like this name:

```
/dev/dasdm1
```

To correlate the disk number with the Linux device name to obtain the `/dev/disk/by-id` name, complete the following steps:

1. Using the example above, issue the `lsdasd` command and scan the output for the line with the device number or the Linux device name:

```
#lsdasd
.
.
0.0.7ab4(ECKD) at ( 94: 48) is dasdm : active at blocksize 4096, 1803060 blocks, 7043 MB
.
```

The number at the start of the entry is the device address, `0.0.7ab4`. The `dasdm` command indicates that the DASD ID is `dasdm`, which is used to identify the device in Linux.

2. List the contents of the directory:

   ```
   ls -l  /dev/disk/by-id
   ```

   The `/dev/disk/by-id` directory displays the name of the same disk, and it links to the Linux name that can be used to identify the disk:

```
lrwxrwxrwx 1 root root   12 2009-07-24 07:46 ccw-IBM.750000000L5611.2bea.b4-part1 -> ../../dasdm1
```

   The name of the file extracted from the output line above (to be used in the `fstab` file) is:

   ```
   "ccw-IBM.750000000L5611.2bea.b4-part1"
   ```

3. Locate this line in the `/etc/fstab` file:

   ```
   dev/dasdm1 /data ocfs2 _netdev,datavolume,nointr  0    0
   ```

4. Replace the line from the previous step with this line:

```
ccw-IBM.750000000L5611.2bea.b4-part1 /datadisk1 ocfs2   _netdev,datavolume,nointr   0     0
```

5. This name will always identify the same disk, and avoid errors.

Another device naming technique

This technique is an alternate device naming technique that applies only to DASD devices.

The example in "Too many DASDs accessible to Linux for read" also shows a strategy to force the different Linux systems on the cluster to use the same name for the same devices. For example, disk device 7c85 is named `/dev/dasdg` on both nodes in this cluster. Towards the end of the line that begins with `parameters=` is the parameter `dasd=` with associated devices:

```
dasd=7435,7436,743a,7535,7536,7c84,7c85,753a,78b4,79b2,7ab2,7c83,7ab4,
7caf,79b3,79b4,7cb0,7cb4,7db6,7eaf,7eb0,7ab6,7137,7239,7339,7539,7bb2
```

On system restart, the DASD devices are named in the order shown, with the `dasd=` parameter, starting with `/dev/dasda` and ending in this case with `/dev/dasdaa`. Modify the DASD parameters of the other servers to have the matching names required. This method applies only to DASD devices.

Too many DASDs accessible to Linux for read

Linux system restart might be made slower due to processing many DASDs. This processing can cause network time-outs and other failures.

There is a problem that might occur when Linux is started on a server in an enterprise environment, in which the system can reach and identify many disks in the storage network. If the access for a system is not limited by a restrictive authentication mechanism, it is possible that when Linux starts, the scanning of all the storage information could take a long time due to the large number of disks.

Linux system restart time in this situation could be hours. The external symptoms can be erratic behavior immediately after system restart, including unsuccessful

network connections and various failures. The problem could stop suddenly when the scanning of the storage information in the background completes.

The problem was resolved on the test system by modifying the `/etc/zipl.conf` file to exclude access to all disks by device number, except those disks that are of interest. The exceptions must also include Fibre Channel Protocol (FCP), cryptographic, and networking devices, as required. If your system is running Linux on z/VM®, the device address from the z/VM console (typically 0009) must be added back, or the system cannot restart.

An important caution is to include all system disks, including back up system disks, because Linux accesses only those disks specified in the `zipl.conf` file (and made active with the `zipl` command before shutdown).

In this example, in the `parameters` section, the `cio_ignore=all` command makes all devices inaccessible to this system. Then the devices are added back individually, by listing them with each device number preceded by an exclamation mark (!). **Bold** is used to illustrate these devices.

IMPORTANT: The list of disks specified with the `cio_ignore=` command must all appear on one continuous line. This list is illustrated with line breaks for the documentation, but in the file it must be specified without line breaks from **cio_ignore** to **TERM=dumb"**. Incorrect configuration information in the `zipl.conf` file might prevent the system from restarting.

```
[SLES_10_SP2]
    image = /boot/image-2.6.16.60-0.42.4-default
    target = /boot/zipl
    ramdisk = /boot/initrd-2.6.16.60-0.42.4-default,0x1000000
    parameters = "root=/dev/disk/by-id/ccw-IBM.750000000L5611.2be4.35-part1
    cio_ignore=all,!7137,!7220,!7239,!7339,!7435,!7436,!743a,!7535,!7536,!7539,
!753a,!7735,!7736,!773a,!78b2,!78b3,!78b4,!79b2,!79b3,!79b4,!7ab2,!7ab4,!7ab6,!
7bb2,!7c82,!7c83,!7c84,!7c85,!7caf,!7cb0,!7cb3,!7cb4,!7db4,!7db6,!7eaf,!7eb0,!7
eb4,!0420,!0421,!0422,!423,!0424,!0425,!0426,!0440,!0441,!0442,!443,!0444,!0445
,!0446,!0600,!0601,!0602,!0610,!0611,!0612,!0620,!0621,!0622,!0630,!0631,!0632,
!0780,!0781,!0782,!0783,!0784,!0785,!0786 dasd=7435,7436,743a,7535,7436,7436,7c
84,7c85,753a,78b4,79b2,7ab2,7c83,7ab4,7caf,79b3,79b4,7cb0,7cb4,7db6,7eaf,7eb0,7
ab6,7137,7239,7339,7539,7bb2 TERM=dumb"
```

After changing the `/etc/zipl.conf` file, enter the `zipl` command to make the change effective for the next system restart.

Bibliography

Support for OCFS2 is provided by both Novell and Oracle, depending on the platform and context.

Novell provides full support for OCFS2 as implemented in SUSE Enterprise Linux Server (SLES). Novell documentation for SLES 10, including OCFS2 usage, commands, and utilities can be found at this Web site:

http://www.novell.com/documentation/sles10/sles_admin/?page=/documentation/sles10/sles_admin/data/b3uxgac.html

Oracle provides full support for implementation in Red Hat Enterprise Linux (RHEL) and Oracle Enterprise Linux, when used with Oracle products. Oracle documentation for OCFS2 on Red Hat Enterprise Linux and EL with Oracle products can be found at these Web sites:

- For an OCFS2 overview, see:
  http://oss.oracle.com/projects/ocfs2/
- For the OCFS2 Version 1.2 FAQ, see:
  http://oss.oracle.com/projects/ocfs2/dist/documentation/v1.2/ocfs2_faq.html
- For the OCFS2 User's Guide (applies to OCFS2 Version 1.2), see:
  http://oss.oracle.com/projects/ocfs2/dist/documentation/v1.2/ocfs2_users_guide.pdf

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 grant 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.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

IBM World Trade Asia Corporation
Licensing 2-31 Roppongi 3-chome, Minato-ku
Tokyo 106-0032, Japan

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.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
Software Interoperability Coordinator, Department 49XA
3605 Highway 52 N
Rochester, MN 55901
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

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.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating

platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. _enter the year or years_. All rights reserved.

If you are viewing this information in softcopy, the photographs and color illustrations may not appear.

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 http://www.ibm.com/legal/copytrade.shtml.

Adobe®, the Adobe logo, PostScript®, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Cell Broadband Engine™ is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Intel®, Intel logo, Intel Inside®, Intel Inside logo, Intel® Centrino®, Intel Centrino logo, Celeron®, Centrino®, Intel® Xeon®, Intel SpeedStep®, Itanium®, Pentium®, and Xeon® are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Java™ and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Terms and conditions

Permissions for the use of these publications is granted subject to the following terms and conditions.

Personal Use: You may reproduce these publications for your personal, noncommercial use provided that all proprietary notices are preserved. You may not distribute, display or make derivative works of these publications, or any portion thereof, without the express consent of the manufacturer.

Commercial Use: You may reproduce, distribute and display these publications solely within your enterprise provided that all proprietary notices are preserved. You may not make derivative works of these publications, or reproduce, distribute

or display these publications or any portion thereof outside your enterprise, without the express consent of the manufacturer.

Except as expressly granted in this permission, no other permissions, licenses or rights are granted, either express or implied, to the publications or any data, software or other intellectual property contained therein.

The manufacturer reserves the right to withdraw the permissions granted herein whenever, in its discretion, the use of the publications is detrimental to its interest or, as determined by the manufacturer, the above instructions are not being properly followed.

You may not download, export or re-export this information except in full compliance with all applicable laws and regulations, including all United States export laws and regulations.

THE MANUFACTURER MAKES NO GUARANTEE ABOUT THE CONTENT OF THESE PUBLICATIONS. THESE PUBLICATIONS ARE PROVIDED "AS-IS" AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE.

Printed in USA