

Storage Options for Oracle Real Application Clusters on Linux

*An Oracle White Paper
August 2004*

Storage Management Options for RAC

Introduction	3
Storage Options	3
Automatic Storage Management	4
Using ASM with Oracle	5
Migrating a Database to or from ASM	6
Advantages	6
Limitations.....	6
Oracle Cluster File System	6
Using OCFS with Oracle	7
Advantages	7
Disadvantages	7
Raw Devices	8
Using Raw Devices with Oracle	8
Advantages	9
Disadvantages	9
Limitations.....	9
Network Attached Storage (NAS)	9
Using NAS with Oracle.....	10
Limitations.....	11

Storage Management Options for RAC

INTRODUCTION

Oracle Real Application Clusters (RAC) is an implementation of Oracle Database where more than one instance, located on different cluster nodes, access the same database. To enable simultaneous access, the database files must be located on shared storage. RAC also requires cluster management software that can provide concurrent access to the shared storage. On Linux systems, Oracle Cluster Ready Services (CRS), performs the cluster management function.

This white paper provides information for DBAs, support representatives, and others about the shared storage options available for Oracle RAC on Linux systems. It does not provide information about installing and configuring the storage hardware and software. For detailed configuration information, see the *Oracle Real Application Clusters Installation and Configuration Guide*.

STORAGE OPTIONS

For Oracle Real Application Clusters 10g, the following storage options are supported on Linux:

- Automatic Storage Management (ASM)
- Oracle Cluster File System (OCFS)
- Raw devices
- Network Attached Storage (NAS)

You must choose one of these storage options to store the following files:

- Database files, including control files, datafiles and redo log files
- The Server parameter file (SPFILE)
- The Oracle Cluster Registry (OCR) file
- The CRS voting disk

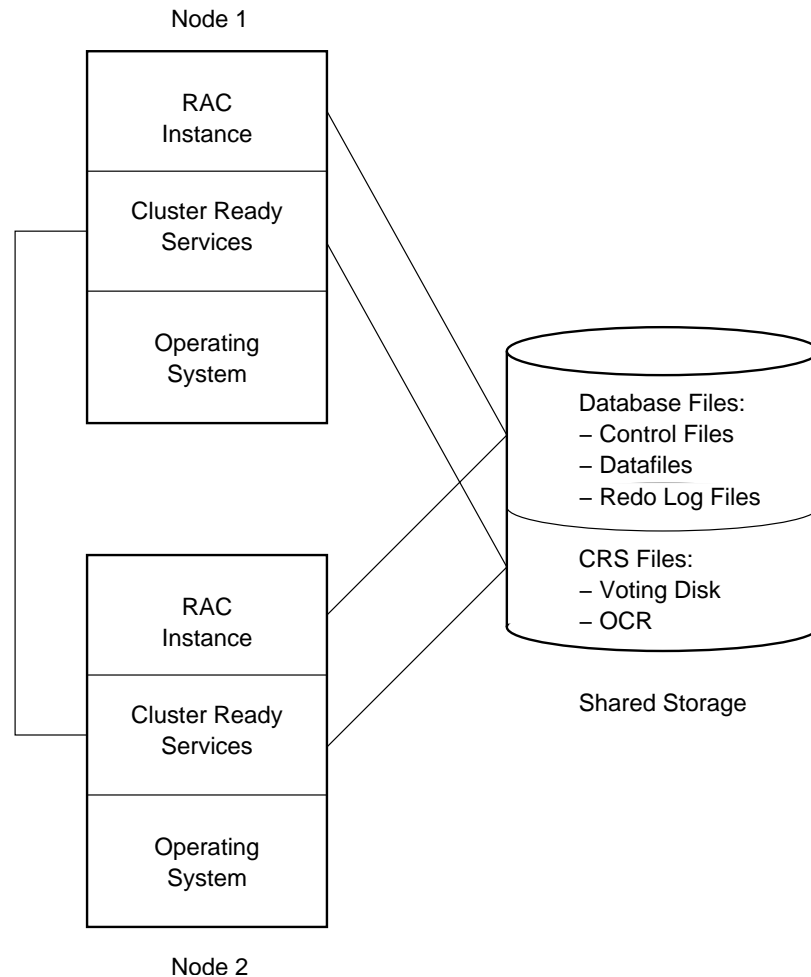
Note: If you choose the NAS storage option, you can also store the RAC and CRS software files (the Oracle home directory and CRS home directory) on a shared file system.

The Oracle Cluster Registry (OCR) contains cluster and database configuration information for RAC and Cluster Ready Services (CRS) such as:

- The cluster node list
- The cluster database instance to node mapping
- The CRS application resource profiles

The voting disk is used by the cluster software for checking the status of the nodes in the cluster.

The following figure shows the architecture of a two node cluster:



Automatic Storage Management

Database storage management has been simplified greatly with the Automatic Storage Management (ASM) feature introduced with Oracle Database 10g. The ASM feature minimizes the DBA efforts in configuring individual database files for the database. ASM is used for database storage only. It provides file system and volume manager capabilities built into the Oracle database kernel. With ASM, you can use either block devices or character devices for database storage. To use block devices you must install and configure the ASM library driver (ASMLib). This library driver enables you to improve the I/O performance of databases that use ASM for storage management. If you intend to install ASM for database storage on

Linux, Oracle recommends that you use the ASM library driver and associated utilities and use them to configure the devices that you want to include in an ASM disk group.

You can use ASM to store the following types of database files:

- Database files
- Control files
- Online redo logs
- Archived redo logs
- Flashback logs
- RMAN backupset pieces
- RMAN Image Copy backups

ASM is not a regular file system. You cannot use it to store the Oracle software files and other components of RAC, for example:

- Trace files
- Alert files
- Audit files
- The Oracle Cluster Registry (OCR)
- The CRS voting disk
- Oracle software files

Using ASM with Oracle

You must configure the ASM disk devices before you create the database. For information about configuring character or block devices for ASM, see the *Oracle Real Application Clusters Installation and Configuration Guide*.

When you select ASM as the storage option in the OUI or DBCA, it lists the disks that you can use and their sizes. The OUI and DBCA use a disk discovery string to identify valid disks. The default disk discovery string on Linux is `/dev/raw/*`. You can change the disk discovery string if required.

Note: If you are using ASMLib, you must choose a custom installation or use DBCA after you have installed the software. You must also specify `ORCL:*` as the disk discovery string.

To configure an ASM disk group during installation or using the DBCA, you must:

1. Specify a disk group name. The default disk group name is DATA.
2. Specify the redundancy level that you want to use for the disk group. The redundancy level specifies how ASM mirrors the contents of the disk group. The following levels are available:

- High (3-way mirroring).
 - Normal (2-way mirroring).
 - External (No ASM mirroring. Use this option if you are using externally mirrored devices such as a RAID 5 device.)
3. If necessary, click **Change Disk Discovery Path** to specify a different disk discovery path.
 4. Select the disks from the list that you want to use for the ASM disk group. As you select the disks, the space requirements are calculated automatically and any additional space requirements are listed.
 5. To create the diskgroup, click **Next**.

The ASM instance is created and named +ASM.

Migrating a Database to or from ASM

You can migrate a database that uses a different storage option to ASM, and you can migrate from ASM to a different storage type. For more information about migrations, see document 252219.1 on the Oracle *Metalink* Web site:

<http://metalink.oracle.com>

Advantages

Simplified database storage management.

Limitations

ASM can be used only for database storage management.

Oracle Cluster File System

Oracle Cluster File System (OCFS) is a shared file system designed specifically for RAC. OCFS eliminates the requirement for Oracle database files to be linked to logical drives.

OCFS enables up to 32 nodes in a cluster to access a shared file system concurrently. Every node sees the same files and data. Compared with the use of raw devices, OCFS simplifies the management of data that needs to be shared between nodes.

With OCFS release 1.0.11, you can use OCFS to store only database files and CRS files. You cannot use it to store other types of files. Support for software files is planned for a future release. This capability will enable you to install Oracle software into a shared file system.

OCFS is currently supported on Red Hat Enterprise Linux 2.1 (x86 and Itanium), Red Hat Enterprise Linux 3 (x86, Itanium, and x86_64), and SUSE Linux Enterprise Server 8 (x86, Itanium, and x86_64) without modifications or patches,

except for any provided by Oracle. Future releases of OCFS are planned to support all Linux distributions and platforms certified to run RAC. For the latest information about OCFS availability and downloads, see the following Web site:

<http://oss.oracle.com/projects/ocfs/>

By using direct disk I/O (O_DIRECT), server performance on OCFS is comparable to performance with raw devices. On Red Hat Enterprise Linux 2.1, Oracle provides direct I/O functions because they are not part of the standard kernel. On other Linux distributions, Oracle uses the native O_DIRECT feature.

You can use the GUI administration utility `ocfstool` to manage OCFS. This utility can list, format, configure, mount, and dismount OCFS volumes. It can also display the volume header, file listing, and list of configured nodes.

Using OCFS with Oracle

When you configure OCFS, make sure that you mount the OCFS file system using the same mount point directory and the appropriate permissions for the `oracle` user on all cluster nodes.

To use OCFS for RAC database file storage, select the File System storage option in the OUI or the Clustered File System storage option in DBCA, then specify the directory where you want to store the database files.

To use OCFS for CRS file storage, specify files names on the OCFS file system when prompted for the location of the OCR and voting disk during the Oracle CRS installation.

For more information about installing and configuring OCFS, refer to the OSS Web site.

Note: In Oracle9i release 2, you cannot use DBCA or OUI to create a database on a shared file system such as OCFS or NAS during the installation. Instead, to create a database on a shared file system, use a command similar to the following to run DBCA after you have installed the software:

```
$ dbca -datafileDestination directory_path
```

Advantages

Using OCFS as a storage option has the following advantages:

- You can use OCFS like a regular file system, which simplifies database file administration, particularly when compared with raw devices.
- There is no limit to the number of files that you can place on an OCFS file system, compared to raw devices, where there is a limit of 255 devices in each cluster.

Disadvantages

Using OCFS as a storage options has the following disadvantages:

- Performance is less than 5% slower than raw devices and typically 2% slower.
- OCFS version 1 does not support regular files, so you cannot place the Oracle software on an OCFS file system.

Raw Devices

A raw device, also known as a raw partition, is a disk partition that is accessed through a character raw device. Raw devices provide a file-like interface through which Oracle can perform reads and writes to the actual disk.

To configure and use raw devices for database creation:

1. Create the required number of partitions with appropriate sizes for each CRS and database file.
2. Identify raw device files that are unused on all cluster nodes.
3. Bind the raw device files to the appropriate partition device files on each cluster node.
4. Set the appropriate owner, group, and permissions on each raw device file on each cluster node.
5. Configure the system to bind the raw devices when the system reboots.

For more information about completing these steps, see the *Oracle Real Application Clusters Installation and Configuration Guide*.

To ensure that the raw devices are bound correctly, follow these steps:

- Check the permissions of the raw device files on every node to ensure that the `oracle` user can read and write to them.
- Verify that the raw devices are bound to the appropriate partition device files on all of the nodes.
- Enter a command similar to the following to determine whether the device is accessible:

```
$ dd if=/dev/raw/raw1 of=/tmp/raw1.txt bs=1024 count=1
```

Using Raw Devices with Oracle

To use raw devices for RAC database file storage, follow these steps:

1. Create a raw device mapping file with entries similar to the following that identifies the raw device file name associated with each database file:

```
system=/dev/raw/raw1
sysaux=/dev/raw/raw2
...
```

See the *Oracle Real Application Clusters Installation and Configuration Guide* for more information about creating this file.

2. Select the Raw Devices storage option in the OUI or DBCA, then specify the path to the raw devices mapping file.

Note: You can also specify the path to this file using the `DBCA_RAW_CONFIG` environment variable before you start OUI.

To use raw devices for storage for the CRS files, specify the device file names for the raw devices when prompted for the location of the OCR and voting disk during the Oracle CRS installation.

Advantages

The raw device storage option provides the highest performance. Raw reads and writes do not use the operating system buffer cache. Raw reads and writes can also move larger buffers than file system I/O.

Disadvantages

Raw devices are difficult to create, administer, and maintain. Their use is recommended for experienced administrators only.

Limitations

You cannot create more than 255 raw devices on Linux (`/dev/raw/raw1` to `/dev/raw/raw255`). On some Linux distributions, you might need to create raw device files with numbers higher than 128.

There are limits to the number of partitions that you can create on a single drive. You can create a maximum of 15 partitions on a SCSI disk or 63 on an IDE disk (using logical partitions).

Network Attached Storage (NAS)

Note: NAS is the only shared storage option currently supported for storing Oracle software files. If you are not using NAS storage, you must use the local file system to store the Oracle software.

You can use a file system on an NAS device to store software files, CRS files, and database files, for example:

- The installation home directories
- Oracle Cluster Registry (OCR)
- The CRS voting disk
- Database files, such as datafiles, control files, redo log files, the server parameter file (SPFILE), and the password file

When exporting the NFS volumes for RAC, make sure that you specify the correct privileges so that the cluster nodes can mount and write to the volumes. The NFS file systems must also be mounted at the same mount point on all cluster nodes. To make sure that the NFS file systems are automatically mounted with the correct options when a node reboots, add an entry similar to the following to the `/etc/fstab` file:

```
filer:/vol/pafiler /install/rac/netapp1 nfs
rw,fg,hard,nointr,vers=3,tcp,rsize=32758,
wsize=32768,noac
```

Oracle recommends the following mount options for RAC installations:

```
rw,fg,hard,nointr,vers=3,tcp,rsize=32768,wsize=32768,noac
```

- The `noac` option ensures that clients in a database cluster maintain a coherent image of the data files on the storage server. You must specify this option.
- The `fg` option ensures that the NFS file systems are available before the database instance starts up.
- The `tcp` option causes TCP to be used, which ensures extra data integrity.
- The `hard` option ensures data integrity in the event of network problems or a cluster failover event.
- The `nointr` option prevents signals from interrupting NFS client operations. Interruptions might occur during a SHUTDOWN ABORT for example, and can cause data corruption.

Implementing and managing an Oracle installation using NAS is relatively simple, as volumes are presented as a regular file system. For more information, refer to your NAS vendor's Web site.

Using NAS with Oracle

When you install Oracle RAC you can specify a directory on an NAS file system as the location for the CRS and Oracle home directories. You install on one node and at the end of the installation the Oracle software is available on all the nodes of the cluster that you chose.

To use NAS for RAC database file storage, select the File System storage option in the OUI or the Clustered File System storage option in DBCA, then specify the directory where you want to store the database files.

To use NAS for CRS file storage, specify files names on the NAS file system when prompted for the location of the OCR and voting disk during the Oracle CRS installation.

Limitations

Only certified NAS devices are supported. For information about certified NAS devices, see the following Web site:

http://www.oracle.com/technology/products/oracle9i/RAC/tech_linux_x86.html

Oracle recommends that you use a dedicated high-performance network interface, for example Gigabit Ethernet, to access the NAS storage.



Storage Options for RAC on Linux
August 2004
Author: Umadevi Byrappa
Contributing Author: Kevin Flood

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
www.oracle.com

Copyright © 2004, Oracle. All rights reserved.

This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission. Oracle is a registered trademark of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.