An Oracle White Paper
December 2009

Oracle Clusterware 11g Release 2 (11.2) – Using standard NFS to support a third voting file for extended cluster configurations

Version 1.2

ORACLE

Introduction

One of the most critical files for Oracle Clusterware are the voting files. With Oracle Clusterware 11g Release 2, a cluster can have multiple (up to 15) voting files to provide redundancy protection from file and storage failures.

The voting file is a small file (500 MB max.) used internally by Oracle Clusterware to resolve split brain scenarios. It can reside on a supported SAN or NAS device. In general, Oracle supports the NFS protocol only with validated and certified network file servers (http://www.oracle.com/technology/deploy/availability/htdocs/vendors_nfs.html)

Oracle does not support standard NFS for any files, with the one specific exception documented in this white paper. This white paper will give Database Administrators a guideline to setup a third voting file using standard NFS.

The first Oracle Clusterware version to support a third voting file mounted using the standard NFS protocol is Oracle Clusterware 10.2.0.2. Support has been enhanced to Oracle Clusterware 11.1.0.6 based on successful tests. Using standard NFS to host the third voting file will remain unsupported for versions prior to Oracle Clusterware 10.2.0.2. All other database files are unsupported on standard NFS. In addition, it is assumed that the number of voting files in the cluster is 3 or more. Support for standard NFS is limited to a single voting file amongst these three or more configured voting files only. This paper will focus on Grid Infrastructure 11.2.0.1.0, since starting with Oracle Clusterware 11.2.0.1.0 Oracle Automatic Storage Management (ASM) can be used to host the voting files.

THE PROCEDURES DESCRIBED IN THIS PAPER ARE CURRENTLY ONLY SUPPORTED ON AIX, HP-UX, LINUX, SOLARIS.

See table1 for a an overview of supported NFS Server and NFS client configurations.

NFS Server	NFS client	Mount option NFS Client	Exports on NFS Server
Linux 2.6 kernel as a minimum requirement	Linux 2.6 kernel as a minimum requirement	rw,bg,hard,intr,rsize=32768,wsize=32768,tcp,noac,vers=3,timeo=600	/votedisk *(rw,sync,all_squash,anonuid=500,anongid=500)
IBM AIX5.3 ML4	IBM AIX5.3 ML4	rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600,vers=3,proto=tcp,noac,sec=sys	/votedisk sec=sys:krb5p:krb5i:krb5:dh:none,rw,access=nfs1:nfs2,root=nfs1:nfs2
Linux 2.6 kernel as a minimum requirement	IBM AIX5.3 ML4 Note 1:	rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600,vers=3,proto=tcp,noac,sec=sys	/votedisk *(rw,sync,all_squash,anonuid=300,anongid=300)
Sun Solaris 10 SPARC	Sun Solaris 10 SPARC	rw,hard,bg,nointr,rsize=32768,wsize=32768,noac,proto=tcp,forcedirectio,vers=3	/etc/dfs/dfstab : share -F nfs -o anon=500 /votedisk
HP-UX 11.31 (minimum requirement – all HP-UX versions prior to 11.31 are unsupported)	HP-UX 11.31 (minimum requirement – all HP-UX versions prior to 11.31 are unsupported)	rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600,noac,forcedirectio 0 0	/etc/dfs/dfstab : share -F nfs -o anon=201 /votedisk
Linux 2.6 kernel as a minimum requirement	HP-UX 11.31 (minimum requirement – all HP-UX versions prior to 11.31 are unsupported)	rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600,noac,forcedirectio 0 0	/votedisk *(rw,sync,all_squash,anonuid=201,anongid=201)
Note 1:	Linux, by default, requires any NFS mount to use a reserved port below 1024. AIX, by default, uses ports above 1024. Use the following command to restrict AIX to the reserved port range: # /usr/sbin/nfso –p -o nfs_use_reserved_ports=1 Without this command the mount will fail with the error: `vmount: Operation not permitted.`		

Table 1: Supported NFS server / client configurations

Extended or Campus Cluster

In Oracle terms, an extended or campus cluster is a two or more node configuration where the nodes are separated in two physical locations. The actual distance between the physical locations, for the purposes of this discussion, is not important.

Voting File Usage

The voting file is used by the Cluster Synchronization Service (CSS) component, which is part of Oracle Clusterware, to resolve network splits, commonly referred to as split brain. A "split brain" in the cluster describes the condition where each side of the split cluster cannot see the nodes on the other side.

The voting files are used as the final arbiter on the status of the configured nodes (either up or down) and are used as the medium to deliver eviction notices. That means, once it has been decided that a particular node must be evicted, it is marked as such in the voting file. If a node does not have access to the majority of the voting files in the cluster, in a way that it can write a disk heartbeat, the node will be evicted from the cluster.

As far as voting files are concerned, a node must be able to access more than the half of the voting files at any time (simple majority). In order to be able to tolerate a failure of *n* voting files, one must have at least 2n+1 configured. (n= number of voting files) for the cluster. Up to 15 voting files are possible, providing protection against 7 simultaneous disk failures. However, it's unlikely that any customer would have enough disk systems with statistically independent failure characteristics that such a configuration is meaningful. At any rate, configuring multiple voting files increases the system's tolerance of disk failures (i.e. increases reliability).

Extended clusters are generally implemented to provide system availability to protect from site failures. The goal is that each site can run independently of the other one when one site fails. The problem in a stretched cluster configuration is that most installations only use two storage systems (one at each site), which means that the site that hosts the majority of the voting files is a potential single point of failure for the entire cluster. If the storage or the site where n+1 voting files are configured fails, the whole cluster will go down, because Oracle Clusterware will loose the majority of voting files.

To prevent a full cluster outage, Oracle will support a third voting file on an inexpensive, low-end standard NFS mounted device somewhere in the network. Oracle recommends putting the NFS voting file on a dedicated server, which belongs to a production environment.

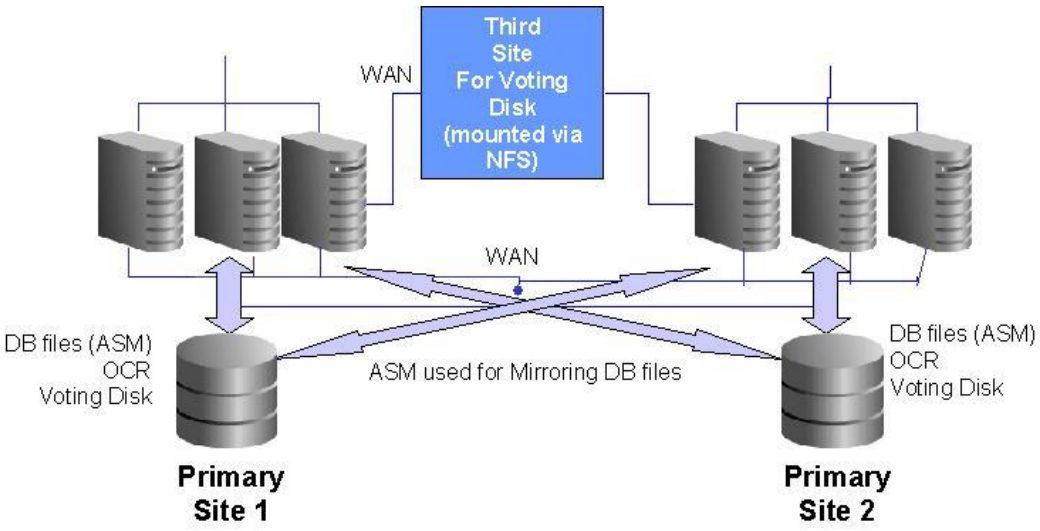


Figure 2: Extended RAC environment with standard NFS Voting file in third site

Voting File Processing

During normal operation, each node writes and reads a disk heartbeat at regular intervals. If the heartbeat cannot complete, the node exits, generally causing a node reboot.

As long as Oracle has enough voting files online, the node can survive. But when the number of offline voting files is greater than or equal to the number of online voting files, the Cluster Communication Service daemon will fail, resulting in a reboot.

The rationale for this is that as long as each node is required to have a majority of voting files online, it is guaranteed that there is one voting file that both nodes in a 2 node pair can see.

Setting up the NFS Server on Linux

THIS CONFIGURATION is CURRENTLY ONLY SUPPORTED ON LINUX
Minimum Kernel version is 2.6

For setting up the NFS server the UID of the software owner and GID of the DBA group are required. The UID and GID should be the same on all the cluster nodes. In order to get the UID and GID of the Oracle user, issue the id command as the Oracle software owner (e.g. oracle) on one of the cluster nodes. To simplify reading this paper we assume that the software owner for the Grid Infrastructure (GI user) and the software owner for the RDBMS part is the same.

```
$ id
uid=500(oracle) gid=500(dba) groups=500(dba)
```

In this case the UID is 500 and the GID is also 500.

As root, create a directory for the voting file on the NFS server and set the ownership of this directory to the UID and the GID obtained for the Oracle software owner:

```
# mkdir /votedisk
# chown 500:500 /votedisk
```

Add this directory to the NFS exports file /etc/exports.
This file should now contain a line like the following:

```
/votedisk *(rw,sync,all_squash,anonuid=500,anongid=500)
```

The anonuid and anongid should contain the UID and GID determined for the GI user and the ASMDBA group on the cluster nodes (here: 500 for UID and 500 for GID)

Check that the NFS server will get started during boot of this server.
For RedHat Linux the respective check is performed as follows:

```
chkconfig --level 345 nfs on
```

Now, start the NFS server process on the NFS server.
For RedHat Linux the respective command is:

```
service nfs start
```

If the new export directory is added to the /etc/exports file while the NFS server process was already running, restart the NFS server or re-export with the command “exportfs –a”.

Check, if the directory containing the voting files is exported correctly by issuing the exportfs –v command as shown below:

```
# exportfs -v
/votedisk
<world>(rw,wdelay,root_squash,all_squash,anonuid=500,anongid=500)
```

Mounting NFS on the Cluster Nodes on Linux

To implement a third voting file on a standard NFS mounted drive, the supported and tested mount options are: rw,bg,hard,intr,rsize=32768,wsize=32768,tcp,noac,vers=3,timeo=600

The minimum Linux kernel version supported for the NFS server is a 2.6 kernel.

To be able to mount the NFS export, create an empty directory as the root user on each cluster node named /voting_disk. Make sure, the NFS export is mounted on the cluster nodes during boot time by adding the following line to the /etc/fstab file on each cluster node:

```
nfs-server01:/votedisk  /voting_disk nfs
rw,bg,hard,intr,rsize=32768,wsize=32768,tcp,noac,vers=3,timeo=600 0 0
```

Mount the NFS export by executing the mount /voting_disk command on each server.

Check, if the NFS export is correctly mounted with the mount command.
This should return a response line as shown below:

```
# mount
nfs-server01:/votedisk on /voting_disk type nfs
(rw,bg,hard,intr,rsize=32768,wsize=32768,tcp,noac,nfsvers=3,
timeo=600,addr=192.168.0.10)
```

Setting up the NFS Server on AIX

For setting up the NFS server the UID of the software owner and GID of the DBA group are required. The UID and GID should be the same on all the cluster nodes. In order to get the UID and GID of the Oracle user, issue the id command as the Oracle software owner (e.g. oracle) on one of the cluster nodes:

```
$ id
uid=500(oracle) gid=500(dba) groups=500(dba)
```

In this case the UID is 500 and the GID is also 500.

As root, create a directory for the voting file on the NFS server and set the ownership of this directory to the UID and the GID obtained for the Oracle software owner:

```
# mkdir /votedisk
# chown 500:500 /votedisk
```

Add this directory to the NFS exports file /etc/exports.
This file should now contain a line like the following:

```
/votedisk -
sec=sys:krb5p:krb5i:krb5:dh:none,rw,access=nfs1:nfs2,root=nfs1:nf
s2
```

Check that the NFS server will get started during boot of this server.
On AIX, check the /etc/inittab for the rcnfs start:

```
# cat /etc/inittab |grep rcnfs
rcnfs:23456789:wait:/etc/rc.nfs > /dev/console 2>&1 # Start NFS
Daemons
```

If the NFS Server is not configured , use smitty to complete the configuration:

```
# smitty nfs
Choose "Network File System (NFS)" ➔"Configure NFS on This
System" ➔ "Start NFS"
Check for "START NFS now, on system restart or both
choose       both    "
```

If the new export directory is added to the /etc/exports file while the NFS server process was already running, restart the NFS server or re-export with the command "exportfs –a".

Check, if the directory containing the voting files is exported correctly by issuing the exportfs –v command as shown below:

```
# exportfs -v
/votedisk -
sec=sys:krb5p:krb5i:krb5:dh:none,rw,access=nfs1:nfs2,root=nfs1:nfs2
```

Mounting NFS on the Cluster Nodes on AIX

To implement a third voting file on a standard NFS mounted drive, the supported and tested mount options are: `rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600, vers=3,proto=tcp,noac,sec=sys`

The minimum AIX version supported for the NFS server is AIX 5L 5.3 ML4 CSP, which includes NFS Server Version 4 (minimum required). All higher versions are also supported.

```
# oslevel -s
5300-04-CSP
```

Use lslpp to determine the exact NFS fileset version:

```
# lslpp -L | grep nfs
  bos.net.nfs.adt     5.3.0.40   C   F    Network File System
  bos.net.nfs.client  5.3.0.44   A   F    Network File System Client
  bos.net.nfs.server  5.3.0.10   C   F    Network File System Server
```

To be able to mount the NFS export, create an empty directory as the root user on each cluster node named /voting_disk. Make sure, the NFS export is mounted on the cluster nodes during boot time by adding the following line to the /etc/filesystems file on each cluster node:

```
/voting_disk:
        dev             = "/votedisk"
        vfs             = nfs
        nodename        = node9
        mount           = true
        options         =
                          rw,bg,hard,intr,rsize=32768,wsize=32768,
                          timeo=600,vers=3,proto=tcp,noac,sec=sys
        account         = false
```

Mount the NFS export by executing the mount /voting_disk command on each server.

Check, if the NFS export is correctly mounted with the mount command.
This should return a response line as shown below:

```
# mount
node9    /votedisk    /voting_disk    nfs3   Nov 03 11:46
rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600,vers=3,proto=tc
p,noac,sec=sys

or use

# lsnfsmnt
/votedisk   node9      /voting_disk          nfs   --
rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600,vers=3,proto=tc
p,noac,sec=sys yes  no
```

Setting up the NFS Server on Solaris

For setting up the NFS server we need to know the UID of the software owner and GID of the DBA group. The UID and GID should also be the same on all the cluster nodes.

To find out the UID and GID issue the id command as the Oracle software owner (e.g. oracle) on one of the cluster nodes,

```
$ id
uid=500(oracle) gid=500(dba) groups=500(dba)
```

In this case the UID is 500 and the GID is also 500.

As root, create the directory for the voting file on the NFS server and set the ownership of this directory to this UID and the GID:

```
# mkdir /votedisk
# chown 500:500 /votedisk
```

Add this directory to the NFS server configuration file /etc/dfs/dfstab. This file should now contain a line like this:

```
share -F nfs -o anon=500 /votedisk
```

The anon= should contain the UID we found for the oracle user on the cluster nodes, in this case 500.

After the entry is added to /etc/dfs/dfstab, you can share the /votedisk directory by either rebooting the system or by using the shareall command:

```
# shareall
```

Check if the votedisk directory is exported correctly by issuing the share command. This command should return a line like this:

```
# share
-               /votedisk   anon=500    ""
```

Unfortunately Solaris only allows to specify an UID for the "anon" option of the "share" command, and it always sets the GID to the same value as the UID if a file gets created by the root user on a NFS client machine. If the UID and GID are not the same the crsctl add css votedisk command afterwards will fail, so it's recommended that the UID and GID are the same.

Mounting NFS on the Cluster Nodes on Solaris

To implement a third voting file on a standard NFS mounted drive, the supported and tested mount options on Solaris 10 are:

```
rw,hard,bg,nointr,rsize=32768,wsize=32768,noac,proto=tcp,forcedirectio,vers=3
```

The minimum Solaris version supported for the NFS server is Solaris 10 SPARC.

To be able to mount the NFS export, as the root user create an empty directory on each cluster node named /voting_disk

Make sure the NFS export is mounted on the cluster nodes during boot time by adding the following line to the /etc/vfstab file on each cluster node;

```
nfs-server01:/votedisk  - /voting_disk  nfs     -       yes
rw,hard,bg,nointr,rsize=32768,wsize=32768,noac,proto=tcp,forcedirectio,vers=3
```

Mount the NFS export by executing the mount /voting_disk command on each server.

Check if the NFS export is correctly mounted with the mount command.

Setting up the NFS Server on HP-UX

The minimum HP-UX version supported for the NFS server HP-UX 11.31. All versions prior to 11.31 are unsupported.

For setting up the NFS server we need to know the UID of the software owner and GID of the DBA group. The UID and GID should also be the same on all the cluster nodes.

To find out the UID and GID issue the id command as the Oracle software owner (e.g. oracle) on one of the cluster nodes,

```
$ id
uid=201(oracle) gid=201(dba)
```

In this case the UID is 201 and the GID is also 201.
As root, create the directory for the voting file on the NFS server and set the ownership of this directory to this UID and the GID:

```
# mkdir /votedisk
# chown 201:201 /votedisk
```

Add this directory to the NFS server configuration file /etc/dfs/dfstab. This file should now contain a line like this:

```
share -F nfs -o anon=201 /votedisk
```

Hereby anon should be set to the UID of the oracle user on the cluster nodes, in this case 201. After the entry is added to /etc/dfs/dfstab, you can share the /votedisk directory by the shareall command:

```
# shareall
```

Check if the votedisk directory is exported correctly by issuing the share command. This command should return a line like this:

```
# share
-               /votedisk   anon=201    ""
```

Unfortunately, HP-UX only allows to specify an UID for the "anon" option of the "share" command, and it always sets the GID to the same value as the UID if a file gets created by the root user on a NFS client machine. If the UID and GID are not the same the crsctl add css votedisk command afterwards will fail, so it's recommended that the UID and GID are the same. Please double check this on the OS used.

Mounting NFS on the Cluster Nodes on HP-UX

To implement a third voting file on a standard NFS mounted drive, the supported and tested mount options on HP-UX 11.31 are:

```
nfs rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600,noac,forcedirectio 0 0
```

To be able to mount the NFS export, as the root user create an empty directory on each cluster node named /voting_disk.
Make sure the NFS export is mounted on the cluster nodes during boot time by adding the following line to the /etc/fstab file on each cluster node;

```
oracle04.bbn.hp.com:/votedisk /voting disk nfs
rw,bg,hard,intr,rsize=32768,wsize=32768,timeo=600,noac,forcedirectio 0 0
```

Mount the NFS export by executing the mount /voting_disk command on each server.

Check if the NFS export is correctly mounted with the mount command.

Adding a 3rd Voting File on NFS to a Cluster using NAS / SAN

Note: Oracle Consulting should be contacted for onsite support, if one is unfamiliar with Oracle Clusterware maintenance and configuration before performing the following steps. In addition, it is strongly recommended to take a backup of the Oracle Cluster Registry (OCR) using *ocrconfig* prior to changing the voting file configuration as suggested.

```
$GRID_HOME/bin/ocrconfig –manualbackup
```

In former releases, backing up the voting disks using the DD-command was a required post-installation task. With Oracle Clusterware 11g Release 2 and later, backing up and restoring a voting disk using the DD command is not supported.

Backing up voting disks manually is no longer required, as voting disks are backed up automatically into the OCR as part of any configuration change. Furthermore, the voting disk data is automatically restored to any added voting disks.

To see which voting files are already configured for the system, use the *$GRID_HOME/bin/crsctl query css votedisk* command. After a default Oracle Clusterware installation, using either a normal redundancy Oracle ASM disk group or using voting disks redundancy based on a NAS device, three voting files should be configured:

```
[root@node1]# crsctl query css votedisk
##  STATE    File Universal Id                File Name Disk group
--  -----    -----------------                --------- ---------
 1. ONLINE   a1625cd6c8b24f0cbfbade52f5e7fa01 (/nas/cluster3/vdsk1) []
 2. ONLINE   0f9d817b97a54f57bf258da457c22997 (/nas/cluster3/vdsk2) []
 3. ONLINE   746a2646226c4fd7bf6975f50b1873a3 (/nas/cluster5/vdsk3) []
```

In the above example, it is assumed that vdsk1 and vdsk2 are on storage in site A and vdsk3 is hosted on a storage in site B. However, what is actually required is a third voting file on storage site C (an NFS mounted device).

Before adding the new voting file, the NFS share must be mounted by adding the mount definition to the /etc/fstab (for Linux) on all of the cluster nodes using the mount options for your platform as described in this paper, followed by issuing an appropriate mount command.

Example for an /etc/fstab entry adding a mount point named /voting_disk used on all nodes:

```
node1:/votedisk /voting_disk nfs
rw,bg,hard,intr,rsize=32768,wsize=32768,tcp,noac,vers=3,timeo=600 0 0
```

After running mount –a on all nodes, perform the following steps as the user root:

- Add the voting file on **one node**:

```
[root@node1 /]# crsctl add css votedisk /voting_disk/vote_3
Now formatting voting disk: /voting_disk/vote_3.
CRS-4603: Successful addition of voting disk /voting_disk/vote_3.
```

Check the ownership for the newly added voting file. If it does not belong to the *id:group* of the oracle owner (e.g. oracle:dba) set the correct ownership using the `chown` command.

To check the new available disk, use the crsctl command again:

```
[root@node1 /]# crsctl query css votedisk
##  STATE    File Universal Id                File Name Disk group
--  -----    -----------------                --------- ---------
 1. ONLINE   a1625cd6c8b24f0cbfbade52f5e7fa01 (/nas/cluster3/vdsk1) []
 2. ONLINE   0f9d817b97a54f57bf258da457c22997 (/nas/cluster3/vdsk2) []
 3. ONLINE   746a2646226c4fd7bf6975f50b1873a3 (/nas/cluster5/vdsk3) []
 4. ONLINE   b09fadba33744fc0bfddb3406553761e (/voting_disk/vote_3) []
Located 4 voting disk(s).
```

Now, there are voting files are on storage side A, B, and C, but still there are two voting files on the storage in site A (vdsk1 and vdsk2). To remove either vdsk1 or vdsk2 use the following command **on only one node**:

```
[root@node1 /]# crsctl delete css votedisk
0f9d817b97a54f57bf258da457c22997
CRS-4611: Successful deletion of voting disk 0f9d817b97a54f57bf258da457c22997.


[root@node1 /]# crsctl query css votedisk
##  STATE    File Universal Id                File Name Disk group
--  -----    -----------------                --------- ---------
 1. ONLINE   a1625cd6c8b24f0cbfbade52f5e7fa01 (/nas/cluster3/vdsk1) []
 2. ONLINE   746a2646226c4fd7bf6975f50b1873a3 (/nas/cluster5/vdsk3) []
 3. ONLINE   b09fadba33744fc0bfddb3406553761e (/voting_disk/vote_3) []
Located 3 voting disk(s).
```

This should be the final configuration: there is now one voting file on storage A (certified storage other), one on storage B (certified storage) and on storage C (NFS mounted). To increase the redundancy, more disks can be added on further storage units. In any case, it needs to be made sure that the majority of the voting files will never be hosted on one storage alone.

The cluster alert log under $CRS_HOME/log/<hostname>/alert<hostname>.log should be monitored during startup in order to see whether the new voting file is actually used. When using the crsctl commands to add or remove a voting file a logfile is written to $GRID_HOME/log/<hostname>/client/crsctl.log

Adding a 3rd Voting File on NFS to a Cluster using Oracle ASM

With Oracle Grid Infrastructure 11g Release 2 the Oracle Cluster Registry (OCR) and the voting files can now be stored in Oracle Automatic Storage Management (ASM) disk groups. While the OCR and the voting files are equally important for the Oracle cluster, this paper will focus on the voting files, since an uninterrupted access to the voting files is vital as described in this paper.

Oracle ASM manages voting disks differently from other files that it stores. When voting files are stored in an Oracle ASM disk group, Oracle Clusterware records exactly where they are located. If Oracle ASM fails, then Cluster Synchronization Services (CSS) can still access the voting files.

Oracle Clusterware requires that voting files are either managed and stored in Oracle ASM completely or not at all. Hybrid configurations, in which some voting files are stored on file systems and some others are stored in Oracle ASM are not supported.

If voting files are stored in ASM, the ASM disk group that hosts the Voting Files will place the appropriate number of voting files in accordance to the redundancy level:

External redundancy: A disk group with external redundancy contains only one voting disk

Normal redundancy: A disk group with normal redundancy contains three voting disks

High redundancy: A disk group with high redundancy contains five voting disks

By default, Oracle ASM puts each voting disk in its own failure group within the disk group and enforces the required number of failure groups and disks in the disk group as listed above.

A failure group is a subset of disks within an Oracle Automatic Storage Management disk group that share a common resource whose failure must be tolerated. Failure groups are used to determine which ASM disks to use for storing redundant copies of data.

For example, four drives that are in a single removable tray of a large JBOD (Just a Bunch of Disks) array are in the same failure group because the tray could be removed, making all four drives fail at the same time.

A quorum failure group is a special type of failure group and disks in these failure groups do not contain user data and are not considered when determining redundancy requirements. For information about failure groups, see Oracle Database Storage Administrator's Guide 11g Release 2 (11.2) "*Oracle ASM Failure Groups" on page 4-24"*.

The COMPATIBLE.ASM disk group compatibility attribute must be set to 11.2 or higher in order to store OCR or voting disk data in a disk group.

The goal of the following steps is to re-configure the existing disk group DATA, currently hosting 3 Voting Files on two storage arrays (as shown in the example output of *crsctl query css votedisk* below), to include a standard NFS based third voting file.

The Oracle Universal Installer (OUI) currently does not support the creation of a quorum failure group or any type of failure group during the initial installation of Oracle Grid Infrastructure.

It is therefore recommended to either configure a very simple disk group initially and then replace this disk group with a disk group that is more suitable for an extended cluster configuration later, or to create a suitable disk group upfront and then modify this disk group following the steps given below. The NFS server and NFS client requirements assumed below are the same as for NAS and a third voting file on NFS.

The following example assumes an existing disk group DATA that currently hosts 3 Voting disks on 2 storage units: /dev/sdg10 is located in storage A and /dev/sdf10 and /dev/sdf11 are located in storage B. It needs to be noted that the fact that those disks are based on different storage units is transparent to Oracle ASM and hence must be enforced by the administrator.

In addition, it is assumed that no failure group assignment, except the default assignment has taken place. This means that each of the disks listed below is placed in a separate failure group per default. As part of the default installation procedure, OUI does not allow a more granular failure group assignment. If a more granular definition is required, it is recommended to create a new Oracle ASM disk group after the initial installation that considers those settings.

```
node1> crsctl query css votedisk
##  STATE    File Universal Id                File Name Disk group
--  -----    -----------------                --------- ---------
 1. ONLINE   3e1836343f534f51bf2a19dff275da59 (/dev/sdg10) [DATA]
 2. ONLINE   138cbee15b394f3ebf57dbfee7cec633 (/dev/sdf11) [DATA]
 3. ONLINE   63514dd105e44fb0bfa741f20126e61c (/dev/sdf10) [DATA]
Located 3 voting disk(s).
```

First, create an empty file on the NFS share before starting the GUI ASMCA command.

```
dd if=/dev/zero of=/voting_disk/vote_stnsp0506 bs=1M count=500

This file must be located on the voting NFS mount point and must
be visible from all nodes before the ASMCA is run.
```

Start the ASMCA (GUI) in order to modify the existing disk group DATA in a way that a quorum failure group containing a standard NFS based voting file can be used for the cluster.

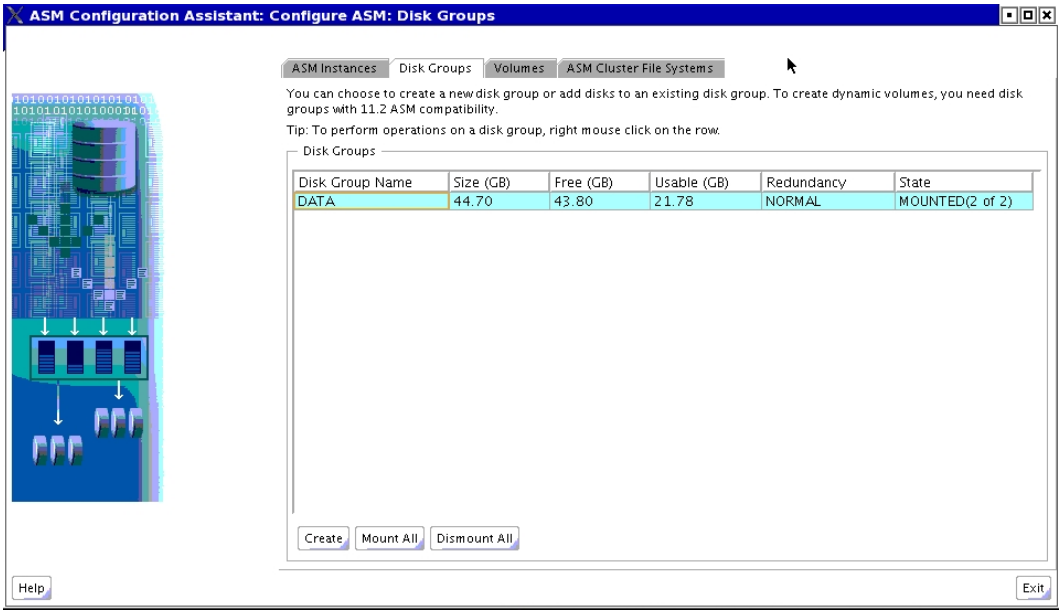


Figure 3: ASMCA GUI

Right click on the existing DATA disk group to expand the window and click on "Add Disks":

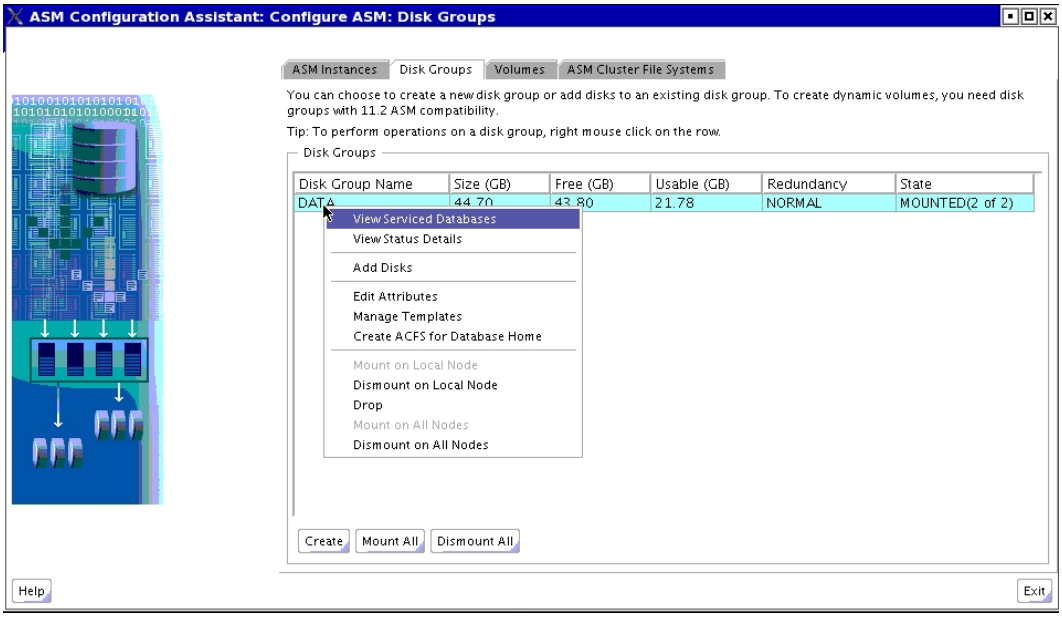


Figure 4: Add Disks to existing disk group

In order to add the NFS based voting file to the ASM disk group that currently only contains disks under /dev/sd*, the ASM disk discovery path needs to be changed in order to contain the NFS mounted file:

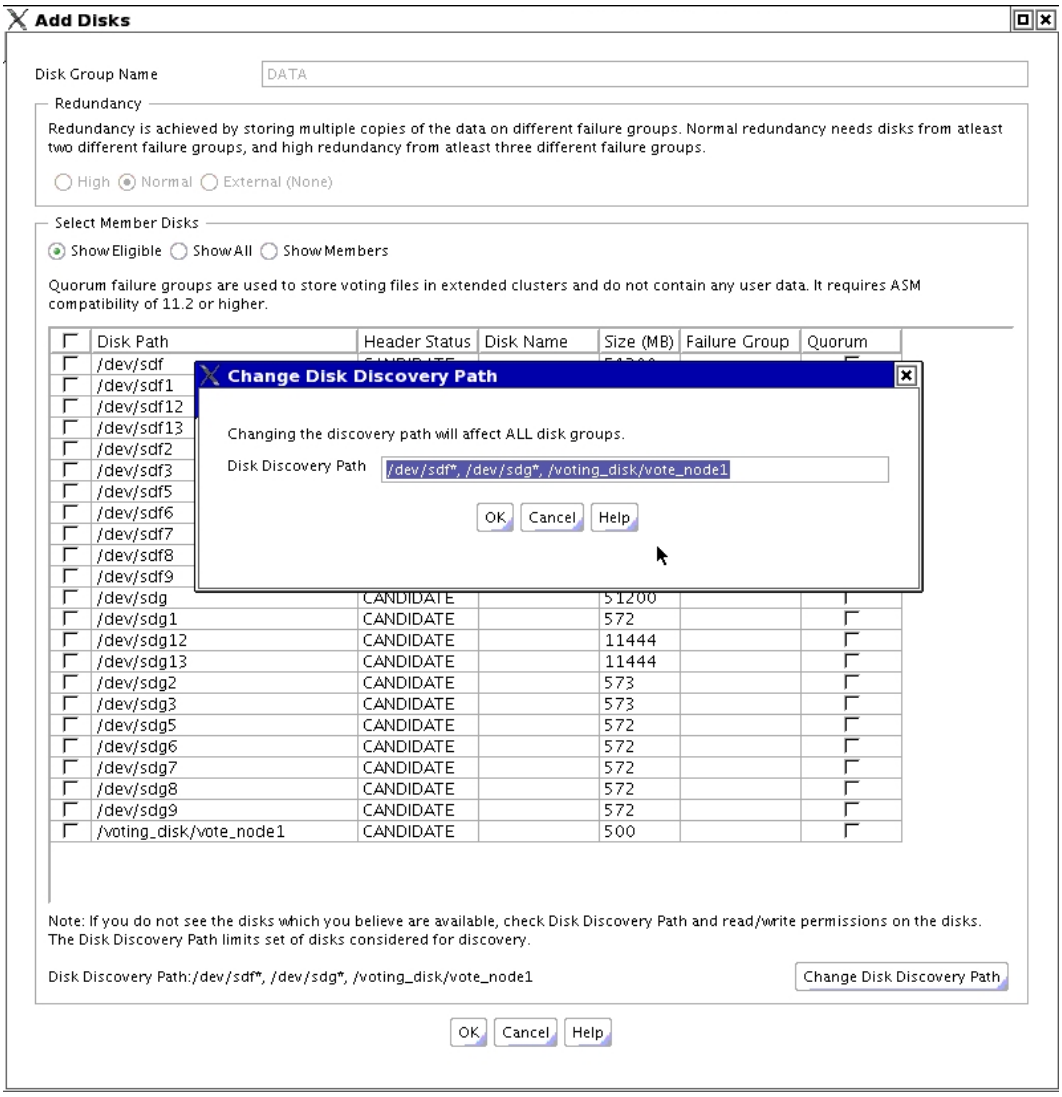


Figure 5: Change the Disk Discovery Path

The next step is to mark the NFS voting disk as the Quorum, then click OK.

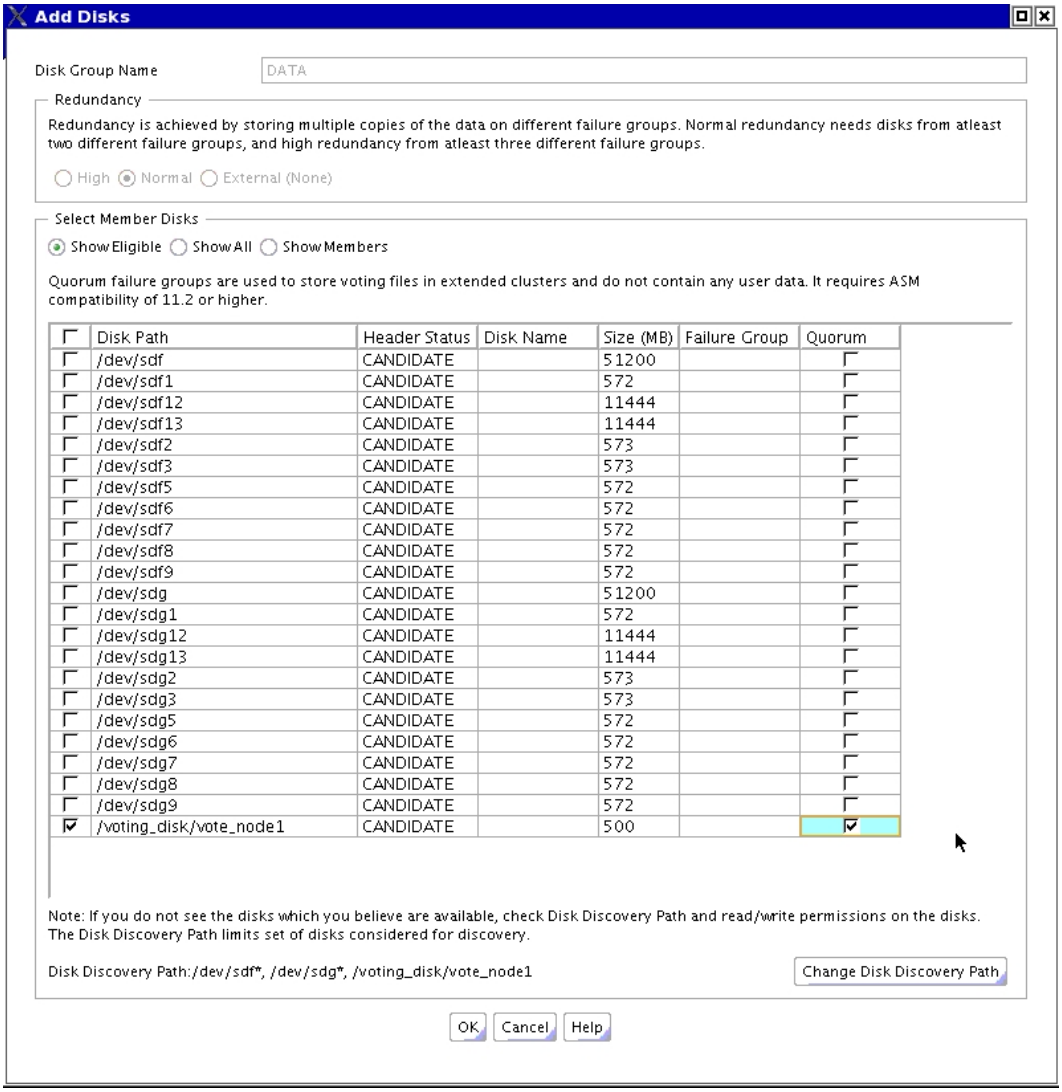


Figure 6: Specify the Quorum disk.

After successful addition exit the ACMCA GUI with exit.

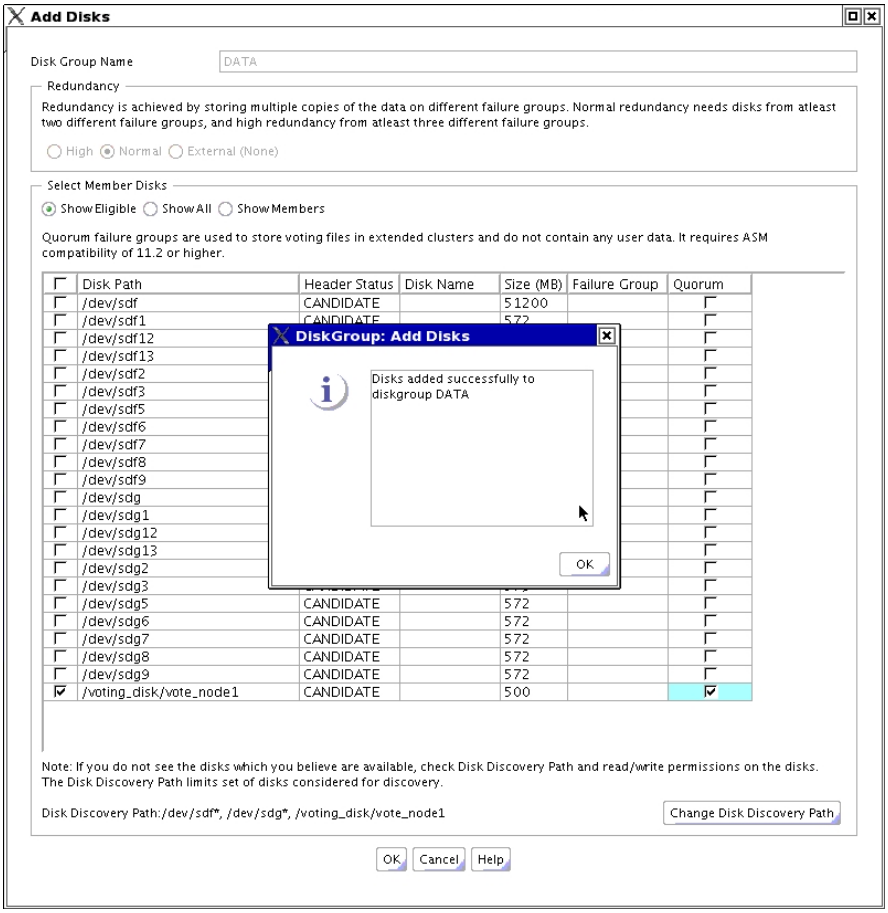


Figure 7: Add Disks screen after successfully adding the Quorum disk.

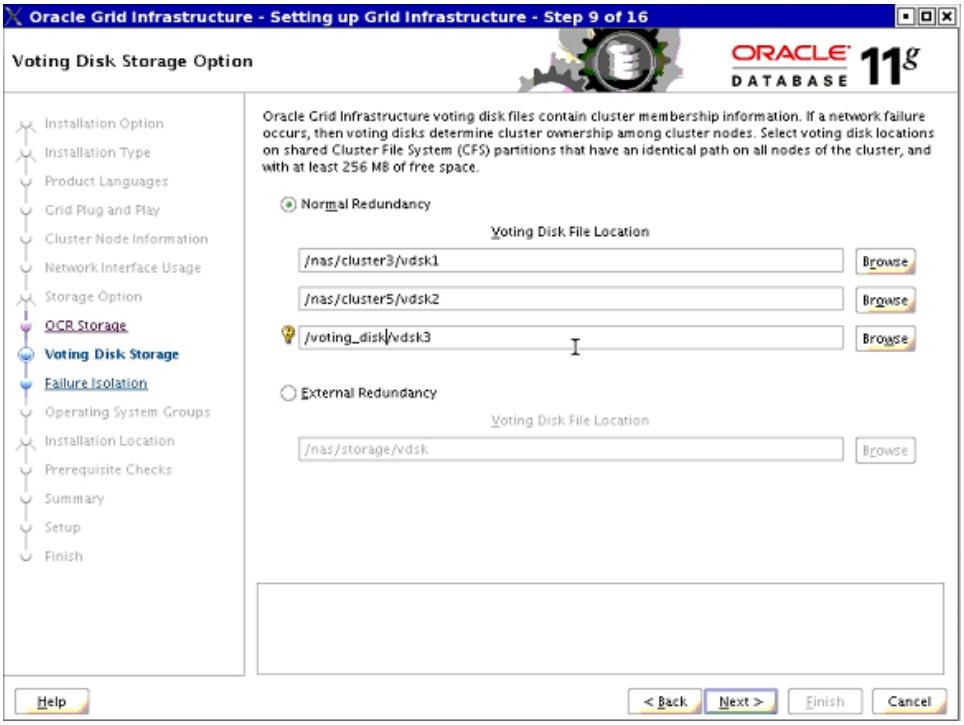
After leaving the ASMCA GUI Oracle Clusterware will automatically notice the disk group change and perform the necessary voting disk relocation. Use the *crsctl query css votedisk* to verify the voting disk re-configuration. The below voting disk configuration is configured in a way that /dev/sdg10 is located on storage A, /dev/sdf11 is located on storage B and /voting_disk/vote_node1 now is located on the external NFS server.

```
node1> crsctl query css votedisk
##  STATE    File Universal Id                File Name Disk group
--  -----    -----------------                --------- ---------
 1. ONLINE   3e1836343f534f51bf2a19dff275da59 (/dev/sdg10) [DATA]
 2. ONLINE   138cbee15b394f3ebf57dbfee7cec633 (/dev/sdf11) [DATA]
 3. ONLINE   462722bd24c94f70bf4d90539c42ad4c (/voting_disk/vote_node1) [DATA]
Located 3 voting disk(s).
```

Configuring 3 File System based Voting Files during Installation

Oracle Clusterware 11g Release 2 supports voting files based on NFS as a configuration option during the installation. In the "Voting Disk Storage Option" screen enter the three different locations in order to use normal redundancy voting files.

The first two locations should be on two different NAS storage boxes, while the third location should be in the third site using standard NFS. The NFS server and NFS client setup must be set up before the Oracle Clusterware installation.



After the installation is complete, check the voting disk and their status as follows:

```
stnsp005_10:19:33>crsctl query css votedisk
##  STATE    File Universal Id                File Name Disk group
--  -----    -----------------                --------- ---------
 1. ONLINE   360041f8fbd94f68bf50b90da7ee02f3 (/nas/cluster3/vdsk1) []
 2. ONLINE   8da8a7c656054f2bbf41f9664d23be52 (/nas/cluster5/vdsk2) []
 3. ONLINE   2a10a4379f334f06bf072b682b462b13 (/voting_disk/vdsk3) []
Located 3 voting disk(s).
```

Known Issues

If the NFS device location is not accessible,

1. Shutting down of Oracle Clusterware from any node using "crsctl stop crs", will stop the stack on that node, but CSS reconfiguration will take longer. The extra time will be equal to the value of css misscount.
2. Starting Oracle Clusterware again with "crsctl start crs" will hang, because some of the old clusterware processes will hang on I/O to the NFS voting file. These processes will not release their allocated resources such as PORT.

These issues are addressed and will be fixed in future versions.

Conclusion: Before stopping or starting the Oracle Clusterware, it should be made sure that the NFS location is accessible using the "df" command for example. If the command does not hang, one may assume that the NFS location is accessible and ready for use.

Appendix A – More Information

For more information on administering refer to the following information:

- Oracle® Database Storage Administrator's Guide 11g Release 2 (11.2)
- Oracle® Clusterware Administration and Deployment Guide 11g Release 2 (11.2)

11g Release 2 (11.2) - Using standard NFS to support a third voting file on a stretch cluster configuration.
December 2009
Version 1.2.1
Author: Roland Knapp
Contributing Authors: Markus Michalewicz

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
oracle.com

Oracle is committed to developing practices and products that help protect the environment

0109