

Using standard NFS to support a third voting disk on a stretch cluster configuration.

An Oracle White Paper
January 2007
Version 1.4

Using standard NFS to support a third voting disk on a stretch cluster configuration.

INTRODUCTION

One of the critical files for the Oracle Clusterware is the voting disk. With Oracle Clusterware 10g Release 2, a cluster can have multiple, up to 32 voting disks, to provide redundancy protection from disk and storage failures. The voting disk is a small file (256 MB max.) on a cluster-aware supported SAN or NAS. In general Oracle supports the NFS protocol only with network file servers.

Oracle does NOT support standard NFS for any files, with the one specific exception documented in this white paper.

Network File System (NFS) is a protocol originally developed by Sun Microsystems in 1984, as a distributed file system which allows a computer to access files over a network as easily as if they were on its local disks.

This white paper will give Database Administrators a guideline to setup a third voting disk using standard NFS.

The first Oracle Database Version to support a third voting disk mounted with standard NFS protocol is Oracle Clusterware 10.2.0.2. On all versions prior to Oracle Clusterware 10.2.0.2 this configuration will remain unsupported. All other database files are unsupported on standard NFS. Additionally, assuming the number of voting disks in the cluster is 3 or more, support for standard NFS is limited to only a single voting disk.

THIS IS CURRENTLY ONLY SUPPORTED ON LINUX, AIX and Solaris.

Please see Figure 1 for a detailed matrix regarding NFS Server and NFS client combination, which are supported right now.

NFS Server	NFS client	Mount option NFS Client	exports NFS Server example
Linux 2.6 kernel as a minimum requirement	Linux 2.6 kernel as a minimum requirement	rw,bg,hard,intr,rsize=32768,wsiz=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,wsiz=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,wsiz=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,wsiz=32768,noac,proto=tcp,forcedirectio,vers=3	/etc/dfs/dfstab : share -F nfs -o anon=500 /votedisk

Figure 1. (The detailed NFS configuration steps see below.)

Note 1:	<p>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:</p> <pre># /usr/sbin/nfso -p -o nfs_use_reserved_ports=1</pre> <p>Without this command the mount will fail with the error:</p> <pre>vmount: Operation not permitted.</pre>
---------	--

STRETCH OR CAMPUS CLUSTER

In Oracle terms, a stretch 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 DISK USAGE

The voting disk is used by the Cluster Synchronization Service (CSS) component of the Oracle Clusterware to resolve network splits, commonly referred to as split brain, where each side of the split cannot see the nodes on the other side. It is used as the final arbiter of the status of configured nodes, either up or down, and to deliver eviction notices, i.e. when a node has been evicted, it is marked as such in the voting disk. If a node does not have access to the majority of the voting disks in the cluster, so that it can write a disk heartbeat, the node will be evicted from the cluster.

As far as voting disks are concerned, a node must be able to access strictly more than half of the voting disks at any time. So if you want to be able to tolerate a failure of n voting disks, you must have at least $2n+1$ configured. ($n=1$ means 3 voting disks). You can configure up to 32 voting disks, providing protection against 15 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 disks increases the system's tolerance of disk failures (i.e. increases reliability).

Stretch clusters are generally implemented to provide system availability in the case where one site has failed. The goal is that each site can run independent of the other when a site failure occurs.

The problem in a stretch cluster configuration is that most of the installations only have two storage systems (1 at each site), which means that the site that houses the majority of the voting disks is a potential single point of failure for the entire cluster.

If the storage or the site where $n+1$ voting disks are configured fails, the whole cluster will go down because we will lose the majority of voting disks.

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

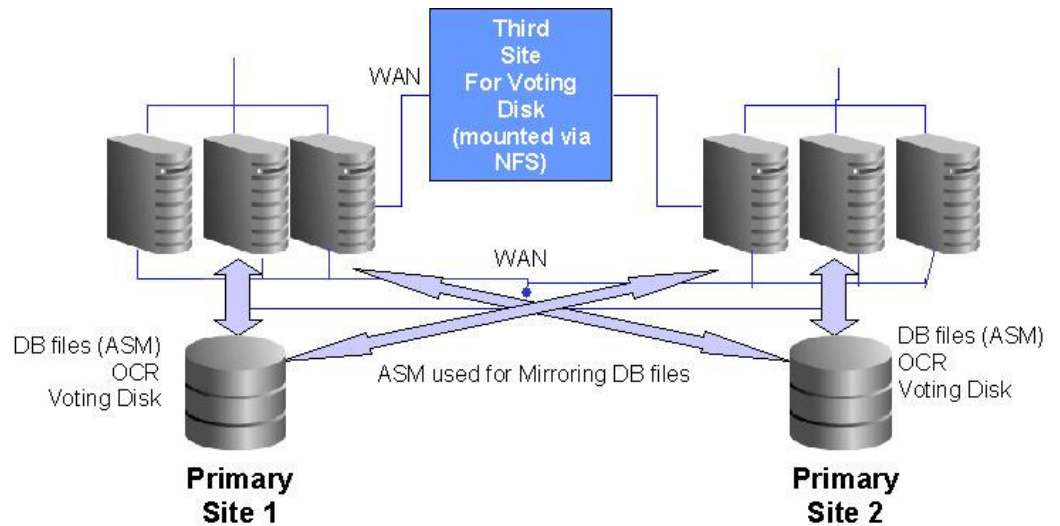


Figure 1 Extended RAC environment using complete Oracle Stack with standard NFS Voting Disk at third site

VOTING DISK PROCESSING

During normal processing, each node writes and reads a disk heartbeat at regular intervals. If the heartbeat can't complete, the node exits, generally causing a node reboot.

As long as Oracle has enough voting disks online, the node can survive, but when the number of offline voting disks is greater than or equal to the number of online voting disks, 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 disks online, there is guaranteed to be one voting disk that both nodes in a 2 node pair can see.

SETTING UP THE NFS SERVER ON LINUX

THIS IS CURRENTLY ONLY SUPPORTED ON LINUX with at minimum a 2.6 kernel.

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 disk 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 exports file /etc/exports. This file should now contain a line like this

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

The anonuid and anongid should contain the UID and GID we found for the oracle user and dba group on the cluster nodes; in this case 500 and 500.

Make sure the NFS server will get started during boot of this server. I.E. for RedHat Linux this could be done like this;

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

Now start the NFS server process on the NFS server. On RedHat Linux this could be done like this;

```
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 votedisk directory is exported correctly by issuing the exportfs -v command. This command should return a line like this;

```
# 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 disk on a standard NFS mounted drive, the supported and tested mount options are,

```
rw,bg,hard,intr,rsize=32768,wsiz=32768,tcp,noac,vers=3,timeo=600
```

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

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;

```
nfs-server01:/votedisk /voting_disk nfs
rw,bg,hard,intr,rsize=32768,wsiz=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 line like this;

```
# mount
nfs-server01:/votedisk on /voting_disk type nfs
(rw,bg,hard,intr,rsize=32768,wsiz=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, we need to create a user and a group, which have same 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 disk 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 exports file /etc/exports. This file should now contain a line like this

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

Make sure the NFS server will get started during boot of this server. Check /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 voting disk directory is exported correctly by issuing the exportfs -v command. This command should return a line like this;

```
# 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 disk on a standard NFS mounted drive, the supported and tested mount options are,

```
rw,bg,hard,intr,rsize=32768,wsiz=32768,timeo=600,vers=3,pr  
oto=tcp,noac,sec=sys
```

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

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

Use `lslpp` to see 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, 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/filesystems` file on each cluster node or use

```
/voting_disk:  
    dev          = "/votedisk"  
    vfs          = nfs  
    nodename     = node9  
    mount        = true  
    options      =  
rw,bg,hard,intr,rsize=32768,wsiz=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 line like this;

```
# mount  
node9  /votedisk  /voting_disk  nfs3  Nov 03 11:46  
rw,bg,hard,intr,rsize=32768,wsiz=32768,timeo=600,vers=3,proto  
=tcp,noac,sec=sys  
or use  
# lsnfsmnt  
/votedisk  node9  /voting_disk  nfs  --  
rw,bg,hard,intr,rsize=32768,wsiz=32768,timeo=600,vers=3,proto  
=tcp,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 disk 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 ""
```

MOUNTING NFS ON THE CLUSTER NODES ON SOLARIS

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

```
rw,hard,bg,nointr,rsize=32768,wsiz=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,wsiz=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.

ADD A THIRD VOTING DISK ON STANDARD NFS TO THE CLUSTER

If you are not familiar with the Oracle Clusterware configuration or Clusterware maintenance please contact Oracle Consulting for on site assistance.

Prior to starting with the voting disk change, it is strongly recommended to backup the Oracle Cluster Repository (OCR) device using ocrconfig,

```
$ORA_CRS_HOME/bin/ocrconfig -export /tmp/ocrbackup -s online
```

To see which voting disks are already configured use the \$ORA_CRS_HOME/bin/crsctl command. Usually after a default Oracle Clusterware installation and using Oracle normal redundancy, there are three voting disks configured,

```
# crsctl query css votedisk
0.      0      /dev/raw/raw3
1.      0      /dev/raw/raw5
2.      0      /dev/raw/raw6
```

In the above case raw3 and raw5 are on storage side A and raw6 is on storage side B, but what is actually required, is a third voting disk on storage side C (the NFS mounted drive).

Before adding the new voting disk, mount the NFS share by adding the mount definition to the (for Linux) /etc/fstab on all of the cluster nodes using the mount options for your platform as described in this paper. For example, an /etc/fstab entry for a mount point named /voting_disk could be used on all nodes:

```
stnsp007.us.oracle.com:/votedisk /voting_disk nfs
rw,bg,hard,intr,rsz=32768,wsz=32768,tcp,noac,vers=3,timeo=
600 0 0
```

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

- Shut down the Oracle Clusterware stack on ALL nodes, as the online addition of voting disks is not supported.

```
# crsctl stop crs
Stopping resources.
Successfully stopped CRS resources
Stopping CSSD.
Shutting down CSS daemon.
Shutdown request successfully issued.
```

- Check on all nodes if the Oracle Clusterware is really down.

```
# crsctl check crs
Failure 1 contacting CSS daemon
Cannot communicate with CRS
Cannot communicate with EVM
```

- Add the voting disk on one node:

```
# crsctl add css votedisk /voting_disk/vote_disk3 -force
Now formatting voting disk:
/voting_disk/vote_disk3
successful addition of votedisk
/voting_disk/vote_disk3
```

Check the ownership for the newly added voting disk. 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,

```
# crsctl query css votedisk
0.      0      /dev/raw/raw3
1.      0      /dev/raw/raw5
2.      0      /dev/raw/raw6
3.      0      /voting_disk/vote_disk3
```

Now we have voting disks on storage side A / B and C but still two on storage side A (raw3 and raw5). To remove the voting on either raw3 or raw5 use the following command on only one node,

```
# crsctl delete css votedisk /dev/raw/raw3 -force
# crsctl query css votedisk
0.      0      /dev/raw/raw5
1.      0      /dev/raw/raw6
2.      0      /voting_disk/vote_disk3
```

This should be the final configuration having a voting disk on storage A (cluster aware), storage B (cluster aware) storage C (NFS mounted). To add more redundancy, more disks can be added on different storage but bear in mind that the majority shouldn't be on one only.

Restart the Oracle Clusterware using `crsctl` on each node,

```
# crsctl start crs
```

Monitor the cluster alert log `$ORA_CRS_HOME/log/<hostname>/alert<hostname>.log` during startup in order to see if the new voting disk is used.

ORACLE CLUSTERWARE INSTALLATION USING A THIRD VOTING DISK ON STANDARD NFS

During the installation of Oracle Clusterware, the Oracle Universal Installer (OUI) will fail to recognize the voting disk location on NFS. When the OUI detects the voting disk is on a shared file system, it will create a dummy file with the name "<some random number>.tmp". After creating this dummy file it will check from all nodes, whether the file is accessible or not. But the problem is, if we use the "noac" option, it will create a fuzzy file, which cannot be verified remotely. So the OUI will declare the 3rd. voting disk location is not sharable.

The workaround is to mount the NFS device without the "noac" mount option, in order to install Oracle Clusterware. Before running root.sh, unmount the NFS device and mount it with "noac" option on all nodes. Then run root.sh.

The best way to add a third voting disk on NFS is to add the disk after Oracle Clusterware is installed, as described above.

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 (default is 60s on Linux and 30s on Unix).
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 disk, these processes will not release their allocated resources such as PORT.
3. If the NFS located voting disk goes offline due to a network failure the cluster alert.log \$ORA_CRS_HOME/log/<hostname>/alert<hostname>.log will not report this. The only file where Oracle tracks this is the \$ORA_CRS_HOME/log/<hostname>/cssd/ocssd.log.
4. If the NFS located voting disk recovers from a network failure this is not reported in either the \$ORA_CRS_HOME/log/<hostname>/alert<hostname>.log or the \$ORA_CRS_HOME/log/<hostname>/cssd/ocssd.log.

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

Conclusion: Before stopping or starting the Oracle Clusterware, the DBA should check if the NFS location is accessible or not. The simplest way to check is to run the "df" command, if it is not hanging, then the NFS location is accessible.

APPENDIX A

For more information on Administering Oracle Clusterware, NFS support and the Voting Disk please refer to the following documents:

Oracle® Database Oracle Clusterware and Oracle Real Application Clusters Administration and Deployment Guide 10g Release 2 (10.2)

Oracle Compatible Network Attached File Servers

http://www.oracle.com/technology/deploy/availability/htdocs/vendors_nfs.html

Note: 294430.1 CSS Timeout Computation in RAC 10g (10g Release 1 and 10g Release 2)

Note.279793.1 - How to Restore a Lost Voting Disk in 10g

Note.268937.1 - Repairing or Restoring an Inconsistent OCR in RAC

Bug: 3972986 MULTIPLE VOTING DISK ADDITION/DELETION/QUERY WORKS OFFLINE BUT NOT ONLINE

Recommended patches: Bug: 5256865 PLACEHOLDER BUG FOR PCW 10.2.0.2 CRS BUNDLE II

http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.commadmn/doc/commadmndita/nfs_intro.htm

ORACLE

Using standard NFS to support a third voting disk on a stretch cluster configuration.

January 2007

Authors: Roland Knapp, Amit Das, Daniel Dibbets

Contributing Authors: John Leys, Barb Lundhild, Tak Wang, Kirk McGowan, Valeria Tatsch, Hernan Saltiel

Version 1.4

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

Copyright © 2006, 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, JD Edwards, PeopleSoft, and Siebel are registered trademarks of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.