



For the Complete Technology & Database Professional



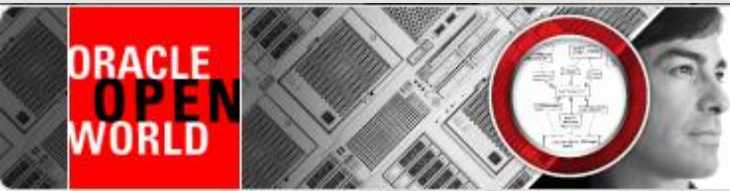
# Ensure the RAC High Availability: Storage and Network Side Story

IOUG Sunday Session #09385

Kai Yu



Dell Oracle Solutions Engineering



**ENGINEERED  
FOR INNOVATION**

**October 2–6, 2011**  
Moscone Center  
San Francisco

## About Author

- **Kai Yu, *kai\_yu@dell.com***

Oracle ACE Director , Senior System Engineer/architect

Dell Oracle Solutions Engineering: [www.dell.com/oracle](http://www.dell.com/oracle)

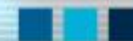
- 16 years Oracle DBA and Oracle Solutions Engineering
- Specialized in Oracle RAC, Oracle E-Business Suite, Oracle VM
- Frequent presenter and whitepaper author
- IOUG Oracle RAC SIG President (2009-2010)
- IOUG Collaborate 09/10/11 Conference committee
- Winner of 2011 OAUG Innovator of the Year Award
- My Oracle Blog: <http://kyuoracleblog.wordpress.com/>

## Dell Global Solutions Engineering Lab



# Agenda

- RAC Database HA Architecture
- Storage Configuration
- Network Configuration
- Clusterware Troubleshooting
- QA



# RAC Database HA Architecture

# RAC Database HA Architecture

## Oracle RAC for Database Fault Tolerance:

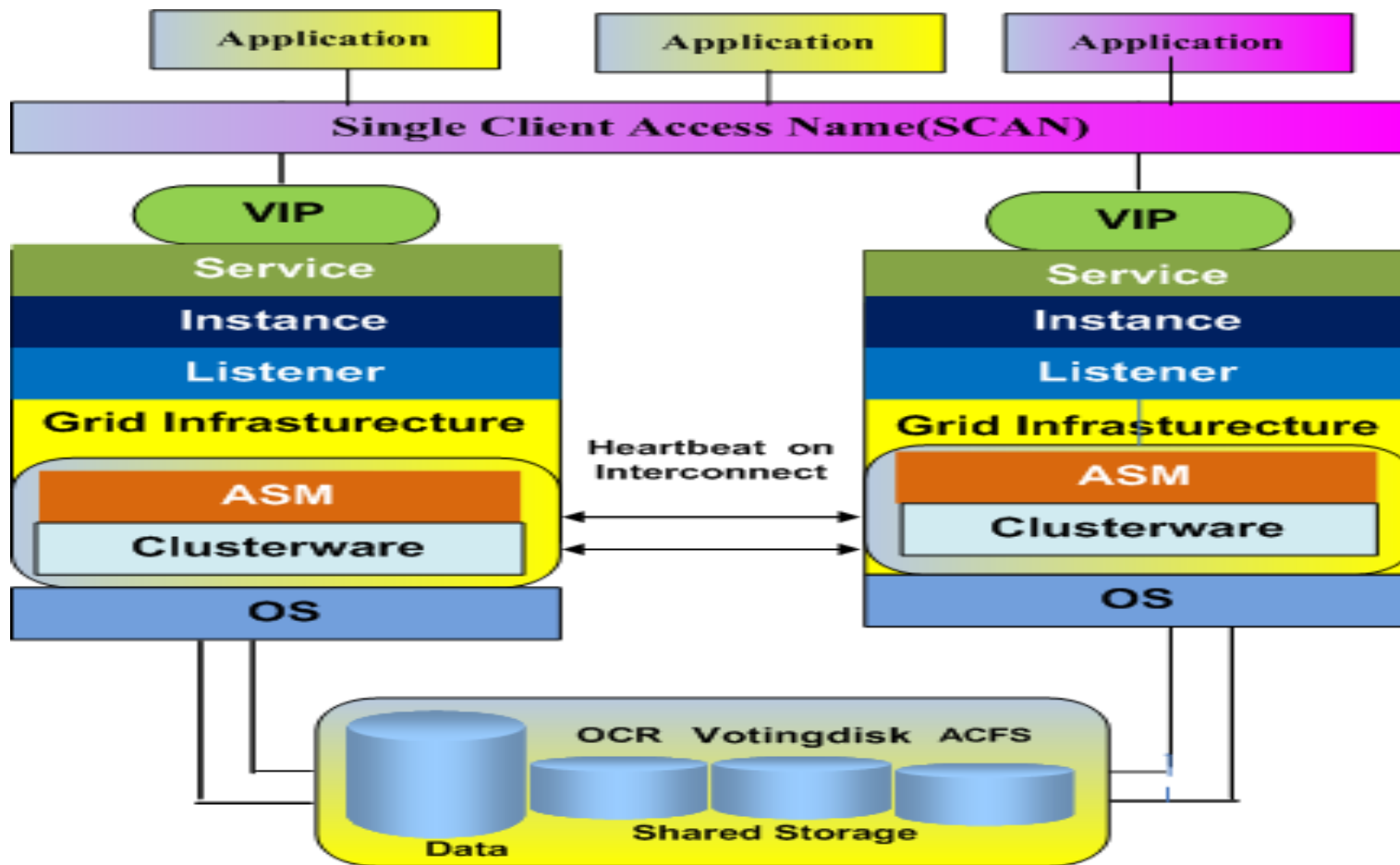
- Redundancy and high availability with n-1 node failure
- Avoid planned or unplanned downtime
- An essential part of Maximal Availability Architecture (MAA)
- Requires fault-tolerant interconnect and shared storage

## Oracle Clusterware for Database High Availability

- Foundation for RAC database and HA features
- Manages RAC database + cluster resources
- Manage failover of Virtual IP to other node
- Restarts failed Oracle processes
- Manages node memberships & prevent the split-brain syndrome

# RAC Database HA Architecture

11g R2 RAC and Clusterware Architecture:

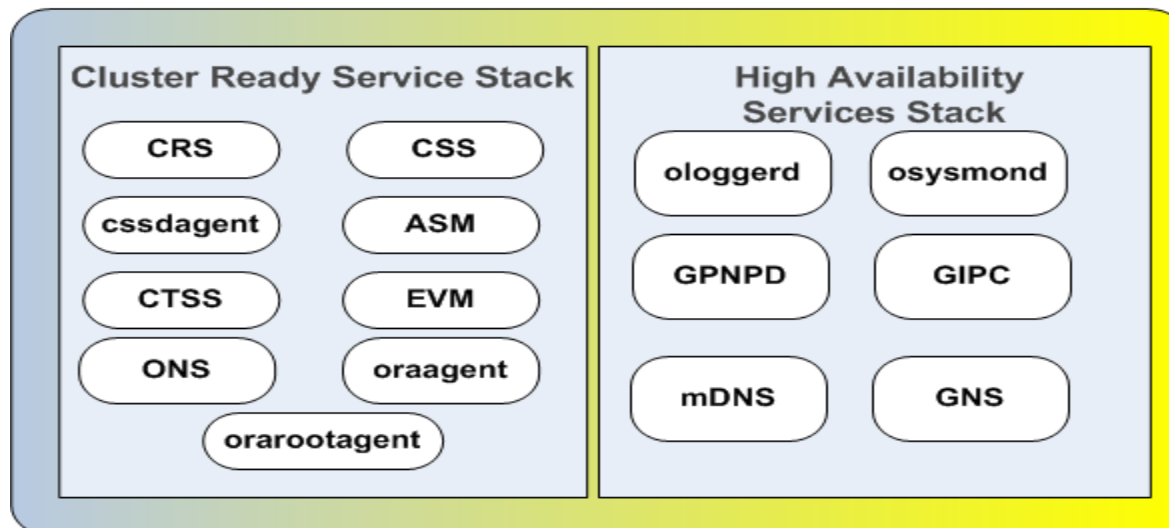


# RAC Database HA Architecture

Grid Infrastructure: Clusterware + ASM

- Two products installed in the same \$GI\_ORACLE\_HOME
- Oracle clusterware components running on each node
- Voting disk: stores cluster membership used by CSS
  - OCR stores information about clusterware resources
  - multiplexed OCR for high availability used by CRS
  - OLR(Oracle Local Registry): \$GI\_HOME/cdata/<host>.olr
  - Oracle Clusterware Stacks

## Oracle 11g R2 Clusterware Stack





# RAC Database HA Architecture

```

grid@k4r815n1:/opt/app/11.2.0/grid
[grid@k4r815n1 grid]$ ps -ef | grep -v grep | grep d.bin
root      /opt/app/11.2.0/grid/bin/ohasd.bin reboot      Oracle HA Service
root      /opt/app/11.2.0/grid/bin/orarootagent.bin      Oracle Agent
root      /opt/app/11.2.0/grid/bin/cssdmonitor           CSS
grid      /opt/app/11.2.0/grid/bin/oraagent.bin          Oracle root Agent
grid      /opt/app/11.2.0/grid/bin/gipcd.bin            GIPC
grid      /opt/app/11.2.0/grid/bin/mdnsd.bin            mDNS
grid      /opt/app/11.2.0/grid/bin/gpnpd.bin           Grid Plug and Play
root      /opt/app/11.2.0/grid/bin/cssdagent           Master DiskMon
grid      /opt/app/11.2.0/grid/bin/diskmon.bin -d
grid      /opt/app/11.2.0/grid/bin/ocssd.bin           CTSS
root      /opt/app/11.2.0/grid/bin/octssd.bin reboot
grid      /opt/app/11.2.0/grid/bin/oclskd.bin           CRS
root      /opt/app/11.2.0/grid/bin/crsd.bin reboot
grid      /opt/app/11.2.0/grid/bin/evmd.bin            Event Monitor
root      /opt/app/11.2.0/grid/bin/oclskd.bin
grid      /opt/app/11.2.0/grid/bin/evmlogger.bin -o /opt/

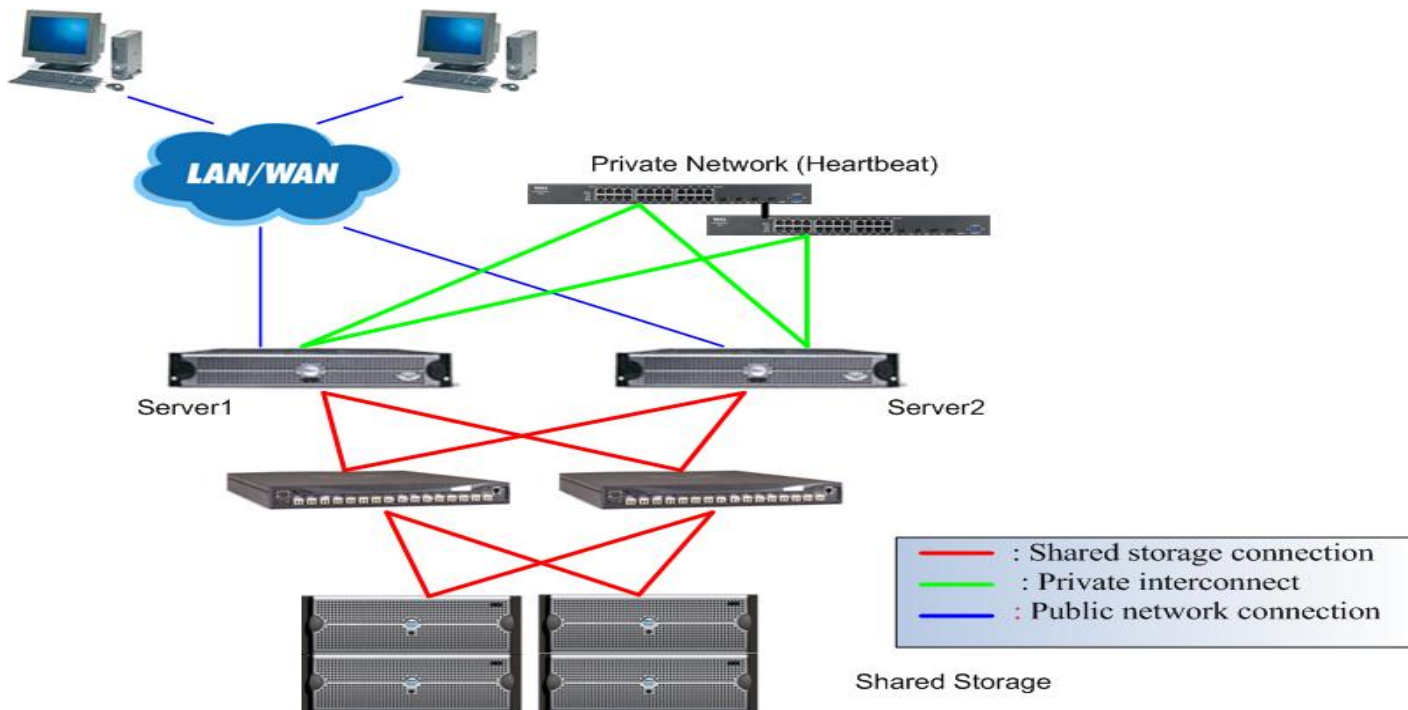
root      /opt/app/11.2.0/grid/bin/orarootagent.bin      Oracle Root Agent
grid      /opt/app/11.2.0/grid/bin/tnslsnr LISTENER -inhe Listener
rit
grid      /opt/app/11.2.0/grid/bin/tnslsnr LISTENER_SCAN1
-inherit
oracle    /opt/app/11.2.0/grid/bin/oraagent.bin          Oracle Agent
oracle    /opt/app/11.2.0/grid/bin/oclskd.bin
[grid@k4r815n1 grid]$
  
```



# Storage Configuration

# Shared Storage Configuration

- Hardware Configuration of Oracle Clusterware
  - Servers, shared storage, interconnect
  - Two interconnect switches for redundant interconnects
  - Butterfly connections to shared storage:  
Servers <-> IO Switches <-> SAN storage



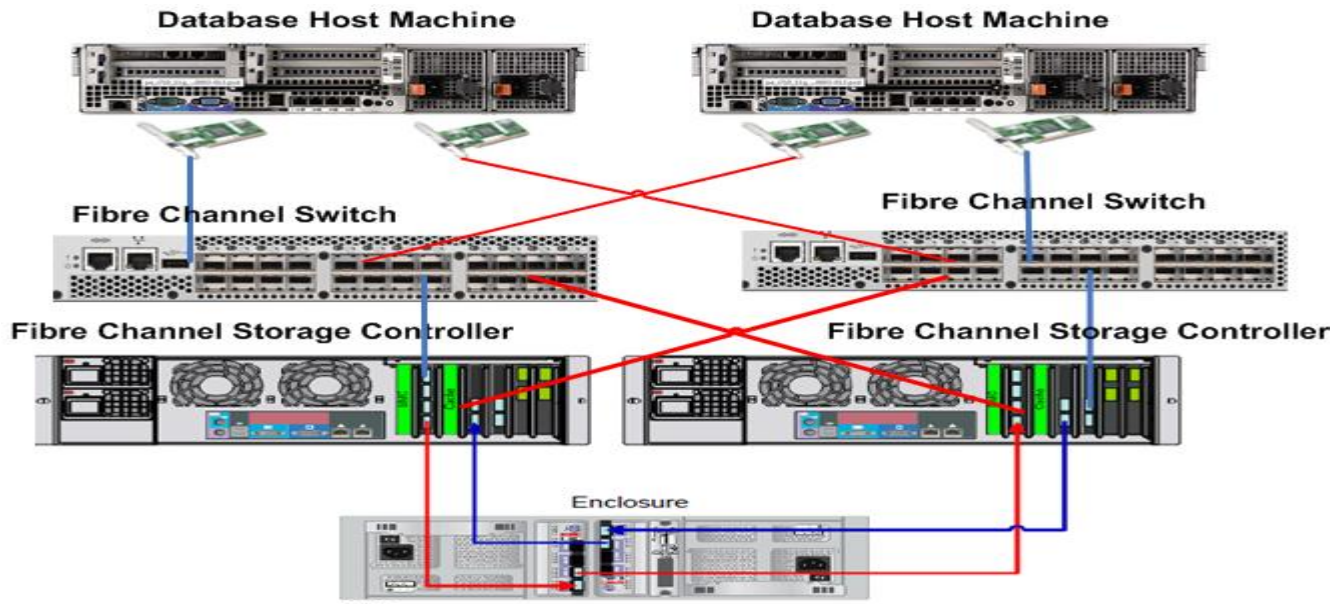
# Shared Storage Configuration

Shared Storage Requirement:

- Shared storage for OCR and votingdisk
- Types: block devices, RAW devices, OCFS2, NFS on NAS
- HA requirement for the shared storage

Physical connections to shared SAN storage

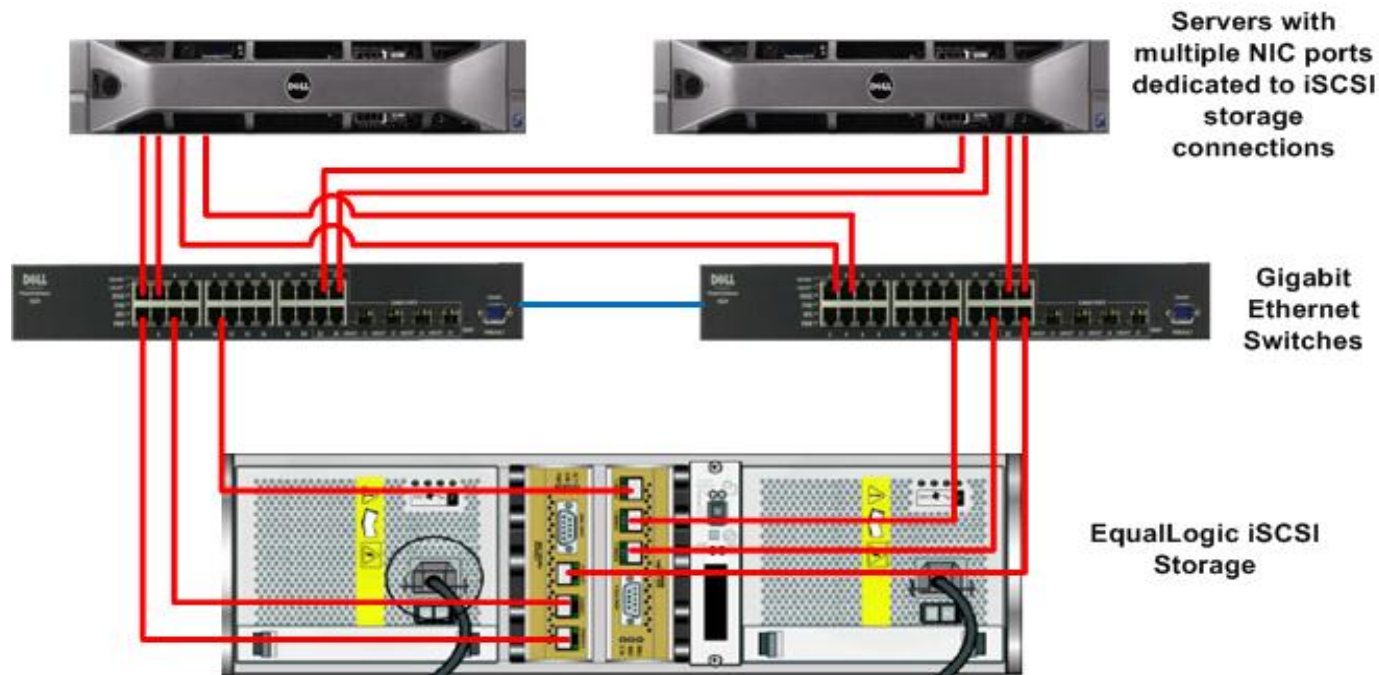
- Fully redundant active-active IO paths: for HA and IO Load balancing: FC storage: dual HBAs and dual FC switches: each server has two independent paths to both storage processors



# Shared Storage Configuration

For iSCSI storage:

Fully redundant IO paths for iSCSI storage, multiple NIC card each server; two Gigabit Ethernet switches. On each storage control module, one network interface connects to one switch and other two network interfaces connects to other switch



# Shared Storage Configuration

## Multipath Devices of the Shared Storage

- Multipathing device driver to combine multiple IO paths
- Example of configuring Linux Device Mapper (DM) for iSCSI
- Verify : `rpm -qa | grep device-mapper`
- Find the unique SCSI ID of the device:

```
$/sbin/scsi_id -gus /block/sdb ←
```

```
36090a028e093fc906099540639aa2149
```

```
$/sbin/scsi_id -gus /block/sde ←
```

```
36090a028e093fc90609954063g9aa2149 ←
```

- Configure multipathing in `/etc/multipath.conf`

```
multipath {  
wwid 36090a028e093fc906099540639aa2149 #<---- for sdb and sde  
  alias votingdisk1  
}
```

- `service multipathd restart`

```
multipath -ll
```

```
ls -lt /dev/mapper/*
```

```
brw-rw---- 1 root disk 253, 8 Feb 18 02:02 /dev/mapper/votingdisk1
```

# Shared Storage Configuration

Example for EMC PowerPath driver for EMC storage:

- Install EMC PowerPath and Naviagent software:  

```
rpm -ivh EMCpower.LINUX-5.3.0.00.00-185.rhel5.x86_64.rpm
```

```
rpm -ivh NaviHostAgent-Linux-64-x86-en_US-6.29.5.0.66-1.x86_64.rpm
```
- Start naviagent agent and PowerPath daemons:  

```
service naviagent start , service PowerPath start
```
- Verify EMC pseudo devices in /proc/partitions:

```
root@k4r815n1:~/rpms
[root@k4r815n1 rpms]# more /proc/partitions | grep emcpowere
 120    64  262144000 emcpowere
 120    65  262140606 emcpowere1
[root@k4r815n1 rpms]# powermt display dev=emcpowere
Pseudo name=emcpowere
CLARiion ID=APM00085001936 [k4r815]
Logical device ID=600601600A9022004C985DB70499DF11 [LUN 5]
state=alive; policy=BasicFailover; priority=0; queued-IOS=0
Owner: default=SP A, current=SP A      Array failover mode: 1
=====
----- Host ----- - Stor - -- I/O Path - -- Stats ---
###  HW Path          I/O Paths   Interf.   Mode   State  Q-IOS  Errors
=====
 6 bfa                ->sdi -> SP B6   active  alive   0      0
 6 bfa                ->sdo -> SP A6   active  alive   0      0
```



# Shared Storage Configuration

- Block Devices vs Raw Devices vs ASM diskgroup
  - RHEL4: 10g: raw devices for 10g, 11gR1: block devices for 11g R2: ASM diskgroup
  - For RHEL 5: raw device service is depreciated:
    - 10g: use udev rules to define the raw devices
    - 11g R1 clusterware: use block devices:  
set the proper ownerships and permissions in /etc/rc.local file  
`chown root:oinstall /dev/mapper/ocr*`  
`chmod 0640 /dev/mapper/ocr*`  
`chown oracle:oinstall /dev/mapper/voting*`  
`chmod 0640 /dev/mapper/voting*`
    - 11g R2 clusterware: Store OCR and Voting disk in ASM
- Configure ASM(Automatic Storage Management)
  - Volume manager and file systems for Oracle database files, OCR, Voting disk, ACFS Cluster file system
  - Benefits: *Provide file systems accessible from more servers*  
*Even distributed IO workload on physical disks*  
*Provide High Availability option*  
*Rebalancing allows to change disk online*  
*Oracle ACFS go beyond of database files*

# Shared Storage Configuration

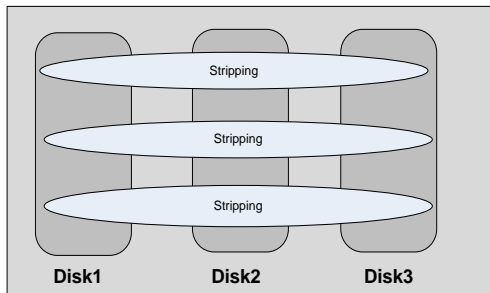
- ASM software: Installed as a part of Grid infrastructure(GI)
- Configure ASMLib and Create ASM disks :
  - Consistent device naming and permission persistency in Linux
  - Provide additional data integrity protection
  - Install the ASMLIB library: Load ASMLib rpms on Linux
  - Built in Oracle UEK kernel. Will not provide ASMLib for RHEL6
  - Configure ASM LIB: *service oracleasm configure*
    - Create ASM disk: *service oracleasm createdisk DATA /dev/emcpowerd1*
    - *\$service oracleasm listdisks*
    - *\$service oracleasm -p querydisks OCR1*
    - Setting ORACLEASM\_SCANORDER, ORACLEASM\_SCANORDEREXCLUDE in /etc/sysconfig/oracleasm for ASMLib to scan the multipath disks and ignore the single path disks, metalink note: #309815.1

```
root@k4r815n1:/dev/oracleasm/disks
[root@k4r815n1 disks]# ls -l /dev/oracleasm/disks
total 0
brw-rw---- 1 grid asmadmin 8, 81 Nov 19 10:35 ACFS
brw-rw---- 1 grid asmadmin 8, 49 Nov 19 10:35 DATA
brw-rw---- 1 grid asmadmin 8, 129 Nov 19 10:35 DATA1
brw-rw---- 1 grid asmadmin 8, 65 Nov 19 10:35 FRA
brw-rw---- 1 grid asmadmin 8, 97 Nov 19 10:35 OCR1
brw-rw---- 1 grid asmadmin 8, 98 Nov 19 10:35 OCR2
brw-rw---- 1 qgrid asmadmin 8, 99 Nov 19 10:35 OCR3
```

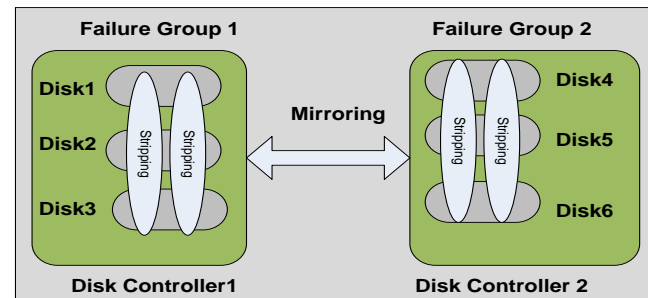
# Shared Storage Configuration

- ASM failure group: mirroring/redundancy level setting:  
External: no ASM mirroring, rely on external redundancy

ASM Disk Group with External Redundancy

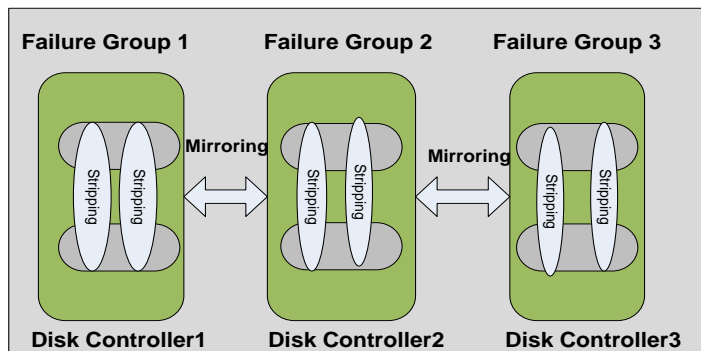


ASM Disk Group with Normal Redundancy



Normal : 2-way mirroring, two failure groups  
High: 3-way mirroring three failures groups.

ASM Disk Group with High Redundancy



```
SQL> create diskgroup new_group HIGH
REDUNDANCY
```

```
failgroup fg1 disk
'ORCL:NEWD1' name newd1,
'ORCL:NEWD2' name newd2
failgroup fg2 disk
'ORCL:NEWD3' name newd3,
'ORCL:NEWD4' name newd4
failgroup fg3 disk
'ORCL:NEWD5' name newd5,
'ORCL:NEWD6' name newd6;
```



# Shared Storage Configuration

## ASM diskgroup for OCR and Voting disks:

- External Redundancy: relay on external RAID configuration
- Normal Redundancy (3 failure groups): 3 voting disk files
- High Redundancy (5 failure groups): 5 voting disk files
- OCR: one copy per diskgroup, up to 5 copies,
- A quorum failure group: only used for OCR and voting disk files

**Create ASM Disk Group**

Select Disk Group Characteristics and select disks

Disk Group Name:

Redundancy:  High  Normal  External

Add Disks

Candidate Disks  All Disks

<input type="checkbox"/>	Disk Path	Size (in MB)	Status
<input type="checkbox"/>	ORCL:ACFS	20479	Candidate
<input type="checkbox"/>	ORCL:DATA	40959	Candidate
<input type="checkbox"/>	ORCL:FRA	20479	Candidate
<input checked="" type="checkbox"/>	ORCL:OCR1	979	Candidate
<input checked="" type="checkbox"/>	ORCL:OCR2	979	Candidate
<input checked="" type="checkbox"/>	ORCL:OCR3	979	Candidate
<input checked="" type="checkbox"/>	ORCL:OCR4	979	Candidate
<input checked="" type="checkbox"/>	ORCL:OCR5	979	Candidate

```

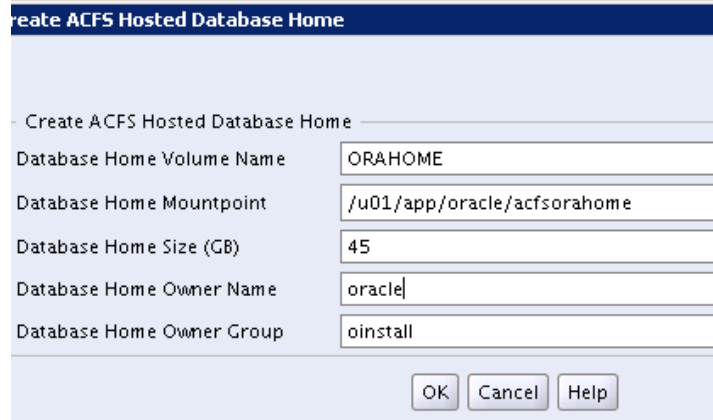
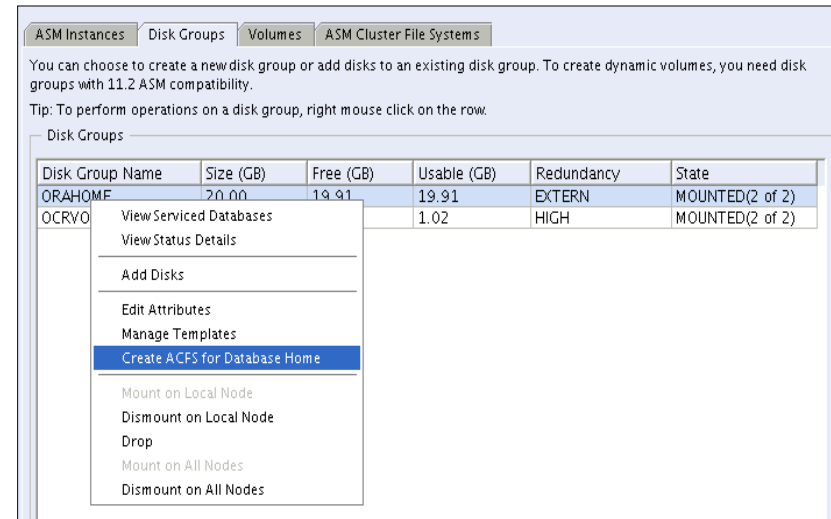
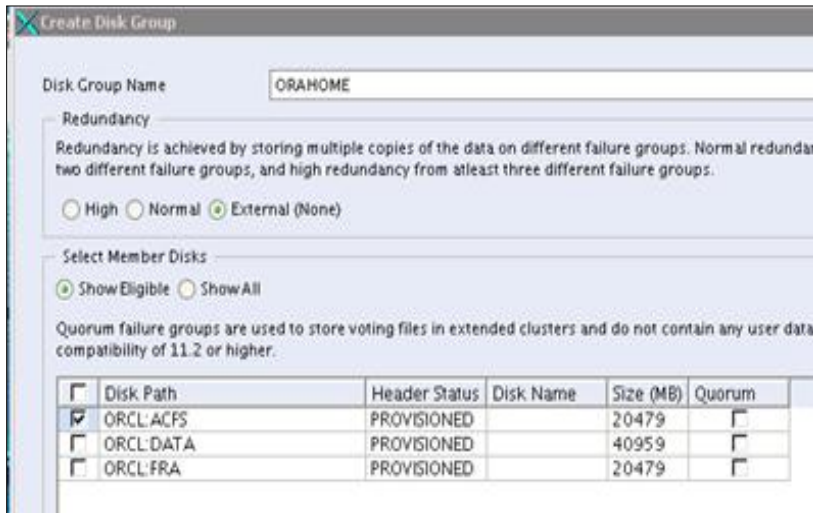
grid@k4r815n1:~
SQL> select path || ' ' || name || ' ' || GROUP_NUMBER || ' ' || FAILGROUP
       2   from v$asm_disk where name like 'OCR%';

PATH||' '||NAME||' '||GROUP_NUMBER||' '||FAILGROUP
-----
ORCL:OCR1 OCR1 4 OCR1
ORCL:OCR2 OCR2 4 OCR2
ORCL:OCR3 OCR3 4 OCR3
ORCL:OCR4 OCR4 4 OCR4
ORCL:OCR5 OCR5 4 OCR5

SQL> !crsctl query css votedisk
## STATE      File Universal Id                        File Name Disk group
-----
 1. ONLINE    6f26b70a85a84f56bf41a1d4d6e87661      (ORCL:OCR5) [OCRVOTDSK]
 2. ONLINE    19546d7b121e4fc6bfc6f224f4fd9de7      (ORCL:OCR4) [OCRVOTDSK]
 3. ONLINE    30e8391c66d54f82bfc5bc8ed239680c      (ORCL:OCR3) [OCRVOTDSK]
 4. ONLINE    06f800a529ad4f02bf45808b04c15683      (ORCL:OCR2) [OCRVOTDSK]
 5. ONLINE    6bc4818a35d44ffe95aaa31831cfb3        (ORCL:OCR1) [OCRVOTDSK]
Located 5 voting disk(s).
    
```

# Shared Storage Configuration

- Create ASM diskgroup for ACFS through asmca:



```
[root@owrac1 app]# df -k
```

```
Filesystem      1K-blocks  Used Available Use% Mounted on
/dev/xvda2      10498192  9755008  635148  94% /
/dev/xvda1       93307    14285   74205  17% /boot
tmpfs           4194304  171116  4023188  5% /dev/shm
/dev/asm/orahome-10 47185920 159172 47026748 1% /u01/app/oracle/acfsorahome
```



# Shared Storage Configuration

Create ASM diskgroups for Database through asmca:

**Create Disk Group**

Disk Group Name:

Redundancy:   
 Redundancy is achieved by storing multiple copies of the data on different failure groups. Normal redundancy uses two different failure groups, and high redundancy from at least three different failure groups.   
 High  Normal  External (None)

Select Member Disks:   
 Show Eligible  Show All

Quorum failure groups are used to store voting files in extended clusters and do not contain any user data. Compatibility of 11.2 or higher.

<input type="checkbox"/>	Disk Path	Header Status	Disk Name	Size (MB)	Quorum
<input checked="" type="checkbox"/>	ORCL:DATA	PROVISIONED		40959	<input type="checkbox"/>
<input type="checkbox"/>	ORCL:FRA	PROVISIONED		20479	<input type="checkbox"/>

**Create Disk Group**

Disk Group Name:

Redundancy:   
 Redundancy is achieved by storing multiple copies of the data on different failure groups. Normal redundancy uses two different failure groups, and high redundancy from at least three different failure groups.   
 High  Normal  External (None)

Select Member Disks:   
 Show Eligible  Show All

Quorum failure groups are used to store voting files in extended clusters and do not contain any user data. Compatibility of 11.2 or higher.

<input type="checkbox"/>	Disk Path	Header Status	Disk Name	Size (MB)	Quorum
<input checked="" type="checkbox"/>	ORCL:FRA	PROVISIONED		512009	<input type="checkbox"/>

```
SQL> Select g.name group_name , d.path disk_path
2 from v$asm_disk d, v$asm_diskgroup g
3 where d.GROUP_NUMBER = g.GROUP_NUMBER;
```

GROUP_NAME	DISK_PATH
ORAHOME	ORCL:ACFS
DATA	ORCL:DATA
FRA	ORCL:FRA
OCRVOTDSK	ORCL:OCR1
OCRVOTDSK	ORCL:OCR2
OCRVOTDSK	ORCL:OCR3
OCRVOTDSK	ORCL:OCR4
OCRVOTDSK	ORCL:OCR5

# Shared Storage Configuration

- ASM Rebalancing and Online Storage Migration
  - Example of performing online storage migration
  - Migrate from DATA1-2 on CX3-80 to DATA3-4 on EqualLogic
  - Steps:
    - Create two volumes DATA3-4, the same size as DATA1-2
    - All database nodes have access to DATA3, DATA4
    - Create two new ASM disks on DATA3, DATA4
    - Replace DATA1, DATA2 with DATA3, DATA4

SQL>

'ORCL:DATA3' name DATA3, 'ORCL:DATA4' name DATA3

drop disk DATA1, DATA2 rebalance power 8;

- Check rebalance status from v\$ASM\_OPERATION

SQL>select GROUP\_NUMBER, SOFAR, EST\_MINUTES

from V\$asm\_operation;

GROUP_NUMBER	SOFAR	EST_MINUTES
--------------	-------	-------------

1	18749	29
---	-------	----

# Network Configuration

# Network Configuration

- *Public Network configuration: virtual IPs*
    - *Virtual IP: failed over to a survived node, no timeout*
    - SCAN (11gR2): a single name of a cluster
      - It has three 3 listeners, 3 virtual IPs for HA and Load balancing
      - A requests first goes to SCAN listener than forwarded to local listener
  - Three ways to configure SCAN VIPs
    - DNS: add three entries to DNS:
    - GNS: assigned through DHCP of GNS(Grid Naming Services)
    - Add into /etc/hosts, only for test environment
  - Set SCAN in DNS:
    - 1) In DNS server: /etc/named/<domain>.zone: add the entries:

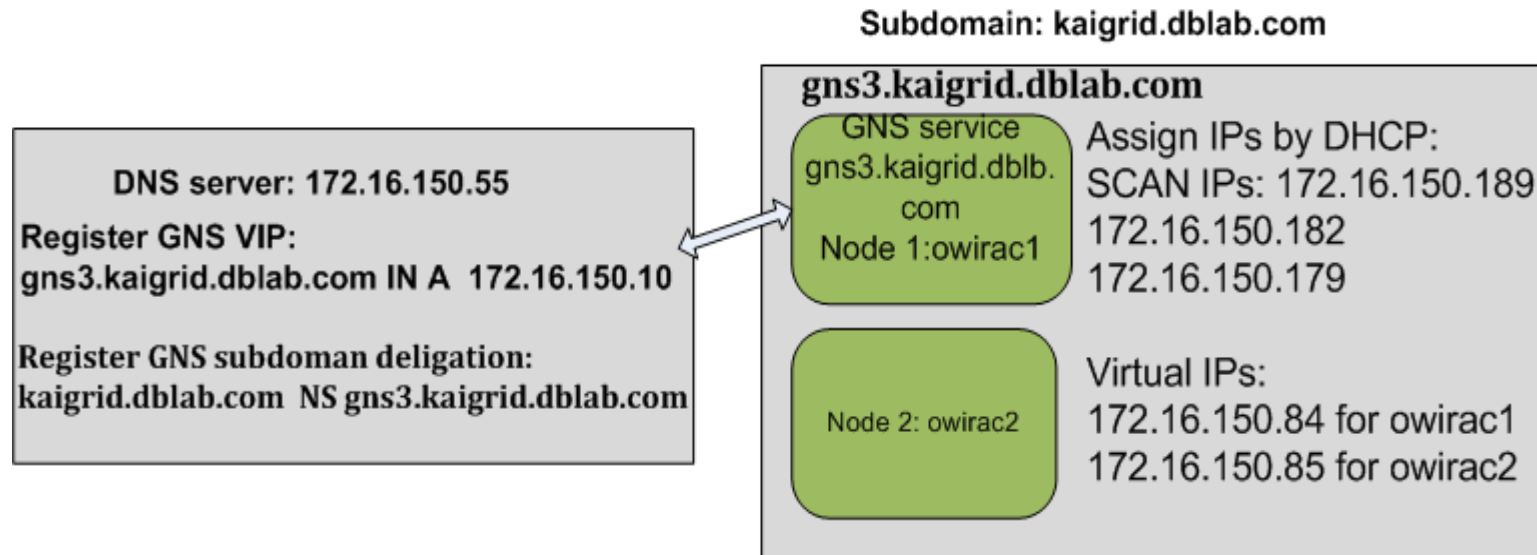
```
km915-scan          IN  A    172.16.150.110
km915-scan          IN  A    172.16.150.111
km915-scan          IN  A    172.16.150.112
km915-N1-VIP IN  A    172.16.150.210
km915-N2-VIP IN  A    172.16.150.211
```
    - 2) One each cluster node: Add to the /etc/resolv.conf:

```
nameserver: 172.16.150.55
search dblab,com
```
- Details: refer to Metalink Support Notes: #1107295.1

# Network Configuration

How GNS service works: (GNS service

- 1) Instead of putting the static VIPs and SCAN IPs in DNS, DNS only lists the static GNS VIP.
- 2) DHCP server provide the IP address for all the VIPs and SCAN IPs
- 3) All dynamic VIPs registered in GNS service.
- 4) DNS forwards every request of the host in the domain to GNS  
xxxx.kaigriddblab.com → gns.kaigriddblab.com



# Network Configuration

How to setup GNS service through 11gR2 clusterware

- On DNS server:
  - Configure GNS VIP in DNS : **kaigrid.dblab.com NS gns3.kaigrid.dblab.com**
  - Configure GNS Subdomain: **gns3.kaigrid.dblab.com IN A 172.16.150.10**
- ON Cluster Nodes: Add entries to /etc/resolv.conf
- Fill out the GNS, SCAN and cluster information on GI installation

```
[root@owirac1 sysconfig]# more /etc/resolv.conf
; generated by /sbin/dhclient-script
;nameserver 10.9.160.254
options attempts: 2
options timeout: 1
nameserver 172.16.150.10
nameserver 172.16.150.55
search kaigrid.dblab.com dblab.com
[root@owirac1 sysconfig]#
```

*GNS Address* (points to 172.16.150.10)

*DNS Address* (points to 172.16.150.55)

*SubDomain for Cluster* (points to kaigrid.dblab.com)



ORACLE 11g DATABASE

Single Client Access Name (SCAN) allows clients to use one name in connection strings to connect to the cluster as a whole. Client connect requests to the SCAN name can be handled by any cluster node.

Cluster Name:

SCAN Name:

SCAN Port:

**Configure GNS**

GNS Sub Domain:   
For example: grid.example.com

GNS VIP Address:

For Details of GNS configuration, Refer to Metalink Support Note # 946452.1



# Network Configuration

## 11gR2 GI(Grid Infrastructure) installation

- Creates GNS service and GNS VIP
- DHCP assigns VIPs/SCAN IPs and registered in GNS

```
[root@owirac1 sysconfig]# srvctl config scan
SCAN name: owirac-cluster-scan.kaigrid.dblab.com, Network: 1/17
SCAN VIP name: scan1, IP: /172.16.150.189/172.16.150.189
SCAN VIP name: scan2, IP: /172.16.150.182/172.16.150.182
SCAN VIP name: scan3, IP: /172.16.150.179/172.16.150.179
```

```
[root@owirac1 sysconfig]# srvctl config vip -n owirac1
VIP exists.:owirac1
VIP exists.: /172.16.150.84/172.16.150.84/255.255.0.0/eth0
```

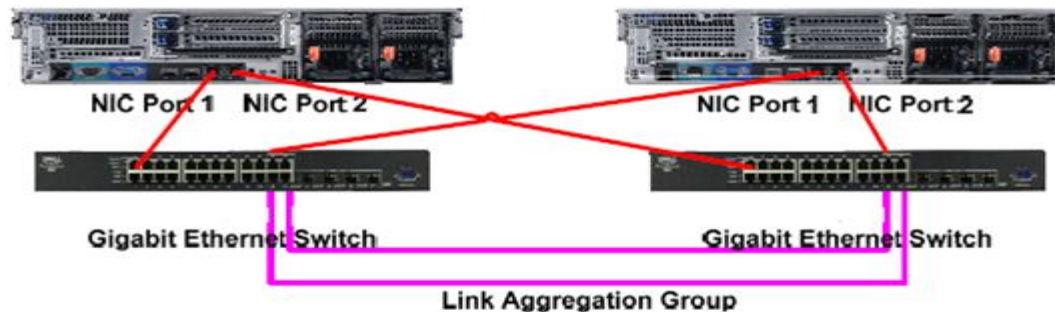
- Creates three SCAN listeners

```
[root@owirac1 ~]# srvctl config scan_listener
SCAN Listener LISTENER_SCAN1 exists. Port: TCP:1521
SCAN Listener LISTENER_SCAN2 exists. Port: TCP:1521
SCAN Listener LISTENER_SCAN3 exists. Port: TCP:1521
```

# Network Configuration

## Private Interconnection Configuration

- Fully Redundant Ethernet Interconnects:  
Two NIC cards, two non-roundtable interconnect switches



- NIC teaming to bond two network interfaces for failover  
DEVICE=eth1      DEVICE=eth2      IPADDR=192.168.9.52  
SLAVE=yes      SLAVE=yes      BOOTPROTO=none
- Configuration best practices from Oracle
  - Separate from the public network
  - Set UDP send/receive buffers to max
  - Use the same interconnect for both Oracle clusterware and Oracle RAC communication



# Network Configuration

- NIC settings for interconnect:
  1. Define control : rx=on, tx=off
  2. Ensure NIC names/slots order identical on all nodes:
  3. Configure interconnect NICs on fastest PCI bus
  4. Compatible switch settings:
    - 802.3ad on NICs = 802.3ad on switch ports
    - MTU=9000 on NICs = MTU=9000 on switch ports
- Recommended minimum Linux kernel configuration for networking  
*net.core.rmem\_default* , *net.core.rmem\_max*, *net.core.wmem\_default*,  
*net.core.wmem\_max*
- Network Heartbeat Misscount: 60 secs for 10g, 30 secs for 11g
- Oprocd, hangcheck replaced by css daemon Agent &monitor in 11gR2
- Redundant Interconnect Usage: on Linux/Unix in 11.2.0.2:
  - use up to four high available virtual IPs (the HAIP)
  - Load balancing and failure protection: automatic relocation.
  - no need of bonding and trunking
- \*\*Known issue with HAIP: if one the private NIC fails, HAIP doesn't fail over the private interconnection to other NIC. it causes node evictions: BUG – 12325672, Doc ID 1323995.1

# Clusterware Troubleshooting



# Clusterware Troubleshooting

- Clusterware utility: crsctl for check, start, and stop ops
  - crsctl -help, crsctl <command\_name> -h
  - Check health of the cluster:

```

grid@k4r815n1:~$ crsctl check crs
CRS-4638: Oracle High Availability Services is online
CRS-4537: Cluster Ready Services is online
CRS-4529: Cluster Synchronization Services is online
CRS-4533: Event Manager is online
[grid@k4r815n1 ~]$ crsctl check cluster -all
*****
k4r815n1:
CRS-4537: Cluster Ready Services is online
CRS-4529: Cluster Synchronization Services is online
CRS-4533: Event Manager is online
*****
k4r815n2:
CRS-4537: Cluster Ready Services is online
CRS-4529: Cluster Synchronization Services is online
CRS-4533: Event Manager is online
*****
    
```

```

grid@k4r815n1:/opt/app/11.2.0/grid/log/k4r815n1/admin
[grid@k4r815n1 k4r815n1]$ pwd
/opt/app/11.2.0/grid/log/k4r815n1
[grid@k4r815n1 k4r815n1]$ ls -lrt | cut -
total
drwxr-x--- grid oinstall diskmon
drwxr-x--- root oinstall ohasd
drwxr-x--- grid oinstall cssd
drwxr-x--- root oinstall crsd
drwxr-x--- grid oinstall gpnpd
drwxr-x--- grid oinstall srvm
-rw-rw-r-- root root alertk4r815n1.log
drwxr-x--- grid oinstall client
drwxr-x--- root oinstall ctssd
drwxr-x--- root oinstall gnsd
drwxrwxr-t grid oinstall racg
drwxr-x--- grid oinstall mdnsd
drwxr-x--- grid oinstall gipcd
drwxr-x--- grid oinstall evmd
drwxrwxr-t root oinstall agent
drwxr-x--- grid oinstall admin
    
```

- Log files and troubleshooting

Log files locations:

*\$GIRD\_Home/log/<host>/alert<host>.log*

*\$GIRD\_Home/log/<host>/<process>/*

# Clusterware Troubleshooting

## Node eviction

- Split brain condition: a node failure partitions the cluster into multiple sub-clusters without knowledge of the existence of others

Possible causes: not responding network heartbeat, disk heartbeat, a hung node or hung ocspd.bin process

- Consequence: data collision and corruption
- IO fencing: fencing the failed node off from all the IOs: STOMITH (Shoot The Other Machine In The Head) algorithm
- Node eviction: pick a cluster node as victim to reboot.  
Always keep the largest cluster possible up, evicted other nodes  
two nodes: keep the lowest number node up and evict other
- Two CSS heartbeats and misscounts to detect node eviction
  1. Network HeartBeat (NHB) : cross the private interconnect  
establish and confirm valid node membership  
CSS misscount: 30 seconds
  2. Disk heartbeat : between the cluster node and voting disk  
CSS misscount, the default is 200 seconds



# Clusterware Troubleshooting

- Node Eviction and Clusterware Reboots:
  - Processes related to reboot
    - OCSSD (CSS daemon): internode health monitoring, monitor a network heartbeat and a disk heartbeat, can evict a node
    - CSSDAGENT: spawning the OCSSD process, monitoring for node
    - CSSDMONITOR: monitors for node hangs, monitors the OCSSD process, monitoring to the OCSSD process for hangs
    - OCSSD (CSS daemon): internode health monitoring, monitor
  - Critical errors in the logfiles in `$GRID_HOME>/log/<nodename>/`
    - Clusterware alert log:
    - cssdagent log(s) in `agent/ohasd/oracssdagent_root`
    - cssdmonitor log(s) in `~agent/ohasd/oracssdmonitor_root`
    - ocssd log(s) in `~cssd`
  - Message files:
    - Linux: `/var/log/messages`
    - Sun: `/var/adm/messages`
    - HP-UX: `/var/adm/syslog/syslog.log`
    - IBM: `/bin/errpt -a > messages.out`

# Clusterware Troubleshooting

- Troubleshooting node eviction
  - Common causes for OCSSD eviction:  
network failure latency exceeds CSS miscount 30 seconds  
Problem with access disk determined by CSS miscount 200 sec  
OCSSD failure,
  - Common causes of : CSSDAGENT OR CSSDMONITOR eviction:  
OS scheduler problem caused by OS locked up in driver or hardware or the heavy loads; thread of CSS demon hung
  - Review the log files, refer to metalink note [1050693.1]

## Node Eviction Diagnosis Case Study

- Case 1 :Node 2 was rebooted in a 2-node 11g R2 cluster on Linux:

OCSSD log:

```
root@k4r815n1:/opt/app/11.2.0/grid/log/k4r815n1/cssd
2010-11-23 17:11:55.221: [    CSSD][1342572864]clssnmPollingThread: node k4r815n2 (2) at 75% heartbeat f
atal, removal in 7.500 seconds
2010-11-23 17:11:59.231: [    CSSD][1353062720]clssnmSendingThread: sending status msg to all nodes
2010-11-23 17:11:59.231: [    CSSD][1353062720]clssnmSendingThread: sent 5 status msgs to all nodes
2010-11-23 17:12:00.232: [    CSSD][1342572864]clssnmPollingThread: node k4r815n2 (2) at 90% heartbeat f
atal, removal in 2.490 seconds, seedhimpd 1
2010-11-23 17:12:02.718: [    CSSD][1342572864]clssnmPollingThread: Removal started for node k4r815n2 (2
), flags 0x3040c, state 3, wt4c 0
2010-11-23 17:12:02.718: [    CSSD][1342572864]clssnmDiscHelper: k4r815n2, node(2) connection failed, en
dp (0x264), probe(0x100000000), ninf->endp 0x264
2010-11-23 17:12:02.718: [    CSSD][1342572864]clssnmDiscHelper: node 2 clean up, endp (0x264), init sta
te 5, cur state 5
```

# Clusterware Troubleshooting

alertk4r815n1.log

```
root@k4r815n1:/opt/app/11.2.0/grid/log/k4r815n1
2010-11-23 17:11:48.206
[cssd(16288)]CRS-1612:Network communication with node k4r815n2 (2) missing for 50% of timeout in
terval. Removal of this node from cluster in 14.510 seconds
2010-11-23 17:11:55.220
[cssd(16288)]CRS-1611:Network communication with node k4r815n2 (2) missing for 75% of timeout in
terval. Removal of this node from cluster in 7.500 seconds
2010-11-23 17:12:00.232
[cssd(16288)]CRS-1610:Network communication with node k4r815n2 (2) missing for 90% of timeout in
terval. Removal of this node from cluster in 2.490 seconds
2010-11-23 17:12:02.719
[cssd(16288)]CRS-1607:Node k4r815n2 is being evicted in cluster incarnation 173918982; details a
t (:CSSNM00007:) in /opt/app/11.2.0/grid/log/k4r815n1/cssd/ocssd.log.
2010-11-23 17:12:11.102
[ohasd(14913)]CRS-8011:reboot advisory message from host: k4r815n2, component: mo163503, with ti
me stamp: L-2010-11-23-17:12:11.052
[ohasd(14913)]CRS-8013:reboot advisory message text: clsnomon_status: need to reboot, unexpected
failure 8 received from CSS
2010-11-23 17:12:33.916
[cssd(16288)]CRS-1601:CSSD Reconfiguration complete. Active nodes are k4r815n1
2010-11-23 17:12:33.926
[crsd(16657)]CRS-5504:Node down event reported for node 'k4r815n2'.
2010-11-23 17:12:37.091
[crsd(16657)]CRS-2773:Server 'k4r815n2' has been removed from pool 'Generic'.
2010-11-23 17:12:37.092
[crsd(16657)]CRS-2773:Server 'k4r815n2' has been removed from pool 'ora.racdb'.
```

Root cause analysis:

A network heartbeat failure triggered node 2 eviction. Found a losing network cable in the single network connection for private interconnect

Solution: Add a redundant network connection with dedicated switch and establishing the network bonding of two network interfaces

- Case: Node 2 was evicted and rebooted due to losing storage connections

```
2010-11-23 18:10:33.162: [    CSSD][1137047872]clssnmvSchedDiskThreads: DiskPingMonitorThread sched
delay 950 > margin 750 cur_ms 2370744 lastalive 2369794
2010-11-23 18:10:35.773: [    CSSD][1357334848]clssnmSendingThread: sending status msg to all nodes
2010-11-23 18:10:35.773: [    CSSD][1357334848]clssnmSendingThread: sent 5 status msgs to all nodes
2010-11-23 18:10:38.689: [    CSSD][1346844992]clssnmPollingThread: node k4r815n1 (1) at 75% heartb
eat fatal, removal in 7.080 seconds
```

```
2010-11-23 18:10:43.699: [    CSSD][1346844992]clssnmPollingThread: node k4r815n1 (1) at 90% heartb
eat fatal, removal in 2.070 seconds, seedhbimpd 1
2010-11-23 18:10:45.766: [    CSSD][1346844992]clssnmPollingThread: Removal started for node k4r815
n1 (1), flags 0x6040e, state 3, wt4c 0
2010-11-23 18:10:45.766: [    CSSD][1346844992]clssnmDiscHelper: k4r815n1, node(1) connection faile
d, endp (0x30c), probe(0x100000000), ninf->endp 0x30c
2010-11-23 18:10:45.766: [    CSSD][1346844992]clssnmDiscHelper: node 1 clean up, endp (0x30c), ini
t state 5, cur state 5
```

```
2010-11-23 18:10:15.898
[cssd(16288)]CRS-1604:CSSD voting file is offline: ORCL:OCR2; details at (:CSSNM00058:) in /opt/app
/11.2.0/grid/log/k4r815n1/cssd/ocssd.log.
2010-11-23 18:10:15.898
[cssd(16288)]CRS-1606:The number of voting files available, 0, is less than the minimum number of v
oting files required, 3, resulting in CSSD termination to ensure data integrity; details at (:CSSNM
00018:) in /opt/app/11.2.0/grid/log/k4r815n1/cssd/ocssd.log
2010-11-23 18:23:52.704
[ohasd(9966)]CRS-2112:The OLR service started on node k4r815n1.
2010-11-23 18:23:53.424
[ohasd(9966)]CRS-8017:location: /etc/oracle/lastgasp has 56 reboot advisory log files, 0 were annou
nced and 0 errors occurred
```

# Clusterware Troubleshooting

- Case 3: Random node evicted in a 11-node 10g R2 cluster on Linux:

/var/log/messages:

11:15:23 racdb7 logger: Oracle clsomon failed with fatal status 12.

Jul 23 11:15:23 racdb7 logger: Oracle CSSD failure 134.

Jul 23 11:15:23 racdb7 logger: Oracle CRS failure. Rebooting for cluster integrity

OCSSD log: \$CRS\_HOME/log/<hostname>/cssd/ocssd.log file

[ CSSD]2008-07-23 11:14:49.150 [1199618400] >WARNING:

clssnmPollingThread: node racdb7 (7) at 50% heartbeat fatal, eviction in 29.720 seconds

..

clssnmPollingThread: node racdb7 (7) at 90% heartbeat fatal, eviction in 0.550 seconds

...

[ CSSD]2008-07-23 11:15:19.079 [1220598112] >TRACE:

clssnmDoSyncUpdate: Terminating node 7, racdb7, misstime(60200) state(3)

Root cause analysis:

- A network heartbeat failure triggered a node eviction on node 7
- Private IP node not pingable right before the node eviction
- Public and private shared a single physical switch
- Solution : Use two dedicated switches for interconnect
- Result: no more node eviction after the switch change



# Clusterware Troubleshooting

## RAC Diagnostic Tools:

- Diagwait: (only need to be reset for pre-11gR2 release)
  - Delay the node reboot for a short time to write all diagnostic messages to the logs.
  - Doesn't increase of probability of data corruption
  - Setup steps: shutdown CRS

```
crsctl set css diagwait 13 -force
```

restart CRS
- Cluster Health Monitor(IPD/OS):
  - A set of tools to collect OS performance data
  - Monitor and record resources degradation and failure related to oracle clusterware and Oracle RAC issue, help to troubleshoot eviction caused by the scheduler issues or high CPU load
  - Historical mode goes back to the time before node eviction
  - Details to refer to Metalink Note [736752.1]
- RAC-RDDT and Oswatcher: Metalink #301138.1, #301137.1
  - Collect information leading up to the time of reboot.
  - From OS utilities: netstat, iostat, vmstat
  - Start: `./startOSW.sh 60 10`

# Thank You and QA

Contact me at [kai\\_yu@dell.com](mailto:kai_yu@dell.com) or visit my Oracle Blog at <http://kyuoracleblog.wordpress.com/>

My OOW 2011

Posted by: [kyuoracleblog](#) | August 13, 2011

☰ Conference

Presentation Schedules

I will present or participate as a panellist of the following OOW sessions:

- 1. Ensure the High Availability and Stability of Oracle RAC: Storage and Network Side Story , Session #09385, 10/2/2011, Sunday, 01:30 PM, Moscone West – 2005*
- 2. Launching the IOUG Virtualization SIG: 360 Degrees of Virtualization for Oracle DBA... , session #28900, IOUG Virtualization panel, 10/2, Sunday, 04:00 PM, Moscone West – 2009*
- 3. Consolidate Oracle E-Business Suite Databases in Oracle Database 11g Release 2 Grid: Case Study, session#08945, 10/4/2011, Tuesday, 10:15 AM, Intercontinental – Intercontinental Ballroom A*
- 4. Configuring and Managing a Private Cloud with Oracle Enterprise Manager , Oracle OpenWorld 2011 session#06980 , 10/4/2011, Tuesday, 05:30 PM, Moscone South – 309, San Francisco*
- 5. Upgrading Oracle Enterprise Manager, Using Best Practices , Oracle OpenWorld 2011 session#0733, 10/6/2011, Thursday, 01:30 PM, Intercontinental – Intercontinental Ballroom A*

I AM A MEMBER OF



**ORACLE**  
ACE Director

**I'm Speaking**



Oct. 2–6, 2011  
Moscone Center  
San Francisco

ARCHIVES

- September 2011
- August 2011
- July 2011
- June 2011
- May 2011
- April 2011
- January 2011
- December 2010
- November 2010



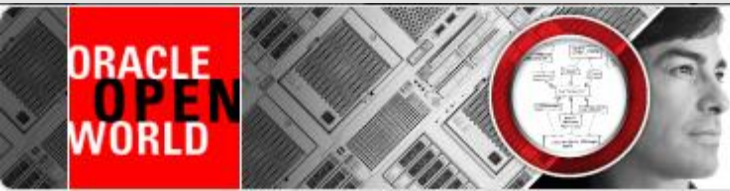
For the Complete Technology & Database Professional



# MARK YOUR CALENDARS!

**COLLABORATE 12**  
**April 22-26, 2012**  
**Mandalay Bay Convention Center**  
**Las Vegas, Nevada**

<http://events.ioug.org/p/cm/ld/fid=15>



**ENGINEERED  
FOR INNOVATION**

**October 2-6, 2011**  
Moscone Center  
San Francisco