

---

# Oracle11g R2 Clusterware : Architecture and Best Practices of Configuration and Troubleshooting



**2011 OTN Latin America Tour**  
**Brazil, Uruguay and Chile, July 2011**

---

Kai Yu  **ORACLE**  
ACE Director  
**Oracle Solutions Engineering**  
**Dell Inc**

# About Author

- Kai Yu, *kai\_yu@dell.com*
  - 16 years with Oracle Technology
  - Focus on Oracle RAC, Oracle VM and Oracle EBS
  - Oracle ACE Director, author and frequent presenter
  - IOUG Oracle RAC SIG President (2009-2010)
  - IOUG Virtualization Committee Member
  - 2011 OAUG Innovator of Year Award Winner
  - Oracle Blog: <http://kyuoracleblog.wordpress.com/>
- Dell Oracle Solutions Engineering: [www.dell.com/oracle](http://www.dell.com/oracle)
  - Oracle Technology Solutions on Dell systems/storages
  - Dell | Oracle Solutions Components
  - Solutions stack: servers, storage, network, OS, virtualization, Oracle RAC, Oracle Applications



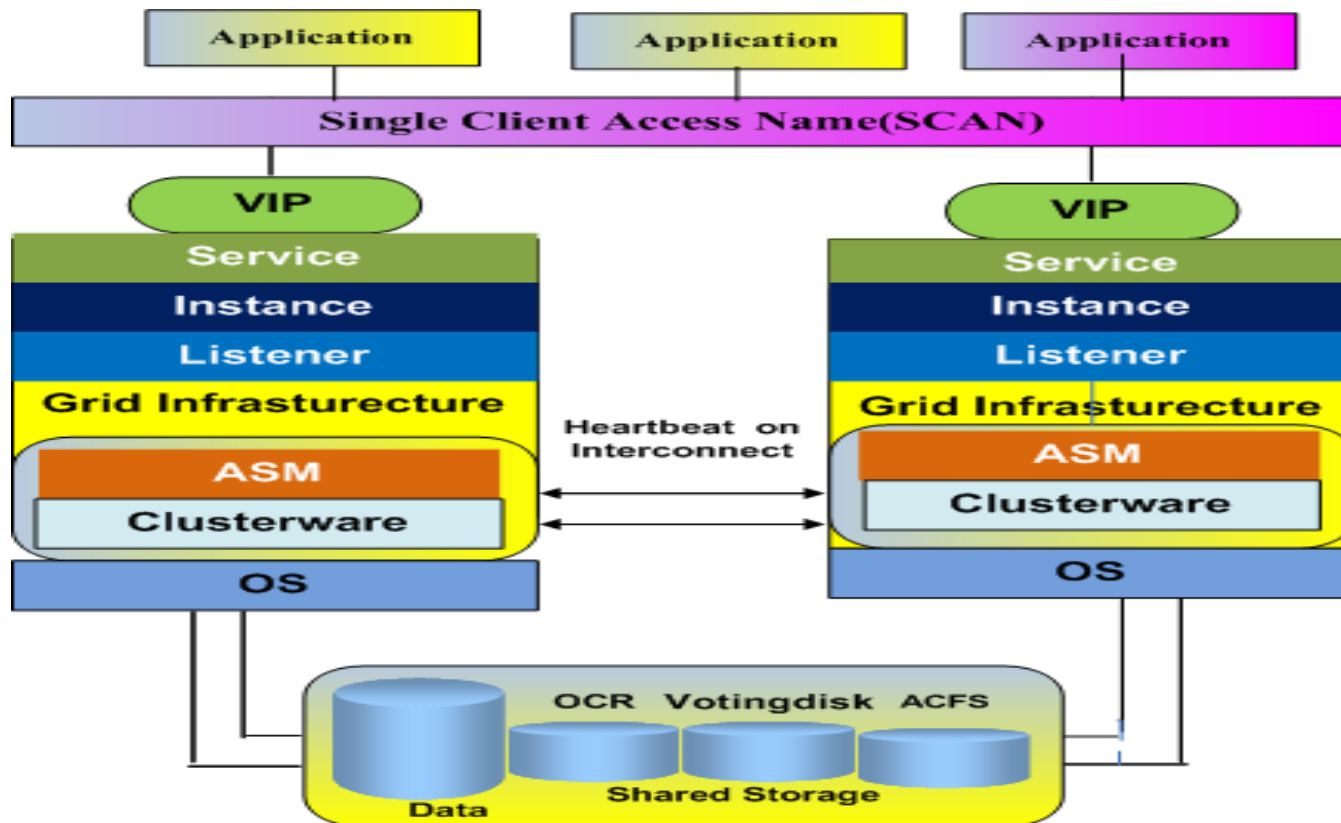
# Agenda

- **Oracle 11g R2 Clusterware Architecture**
- Storage Configuration
- Network Configuration
- Managing Oracle Clusterware
- Clusterware Troubleshooting
- RAC Database Case Study
- QA



# Oracle Clusterware Architecture

- Oracle Clusterware:
  - Role : Provide the infrastructure to run RAC
  - Allow cluster nodes to communicate each other

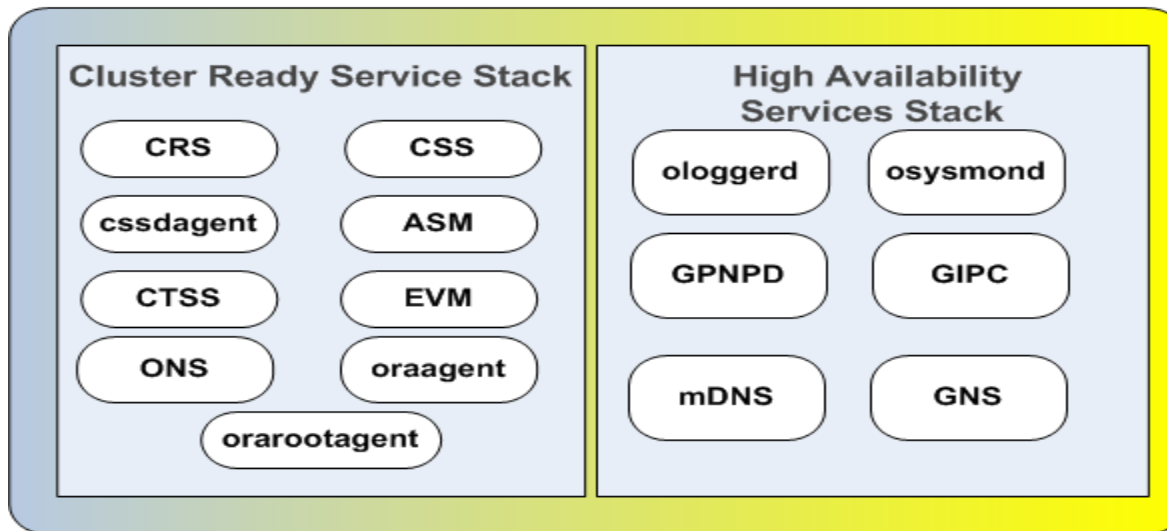


11g R2 Clusterware and RAC

# Oracle Clusterware 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



# Oracle Clusterware 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 Oracle root Agent
grid      /opt/app/11.2.0/grid/bin/oraagent.bin Oracle 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 CSS 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 CRS
grid      /opt/app/11.2.0/grid/bin/oclskd.bin
root      /opt/app/11.2.0/grid/bin/crsd.bin reboot Event Monitor
grid      /opt/app/11.2.0/grid/bin/evmd.bin
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]#
```

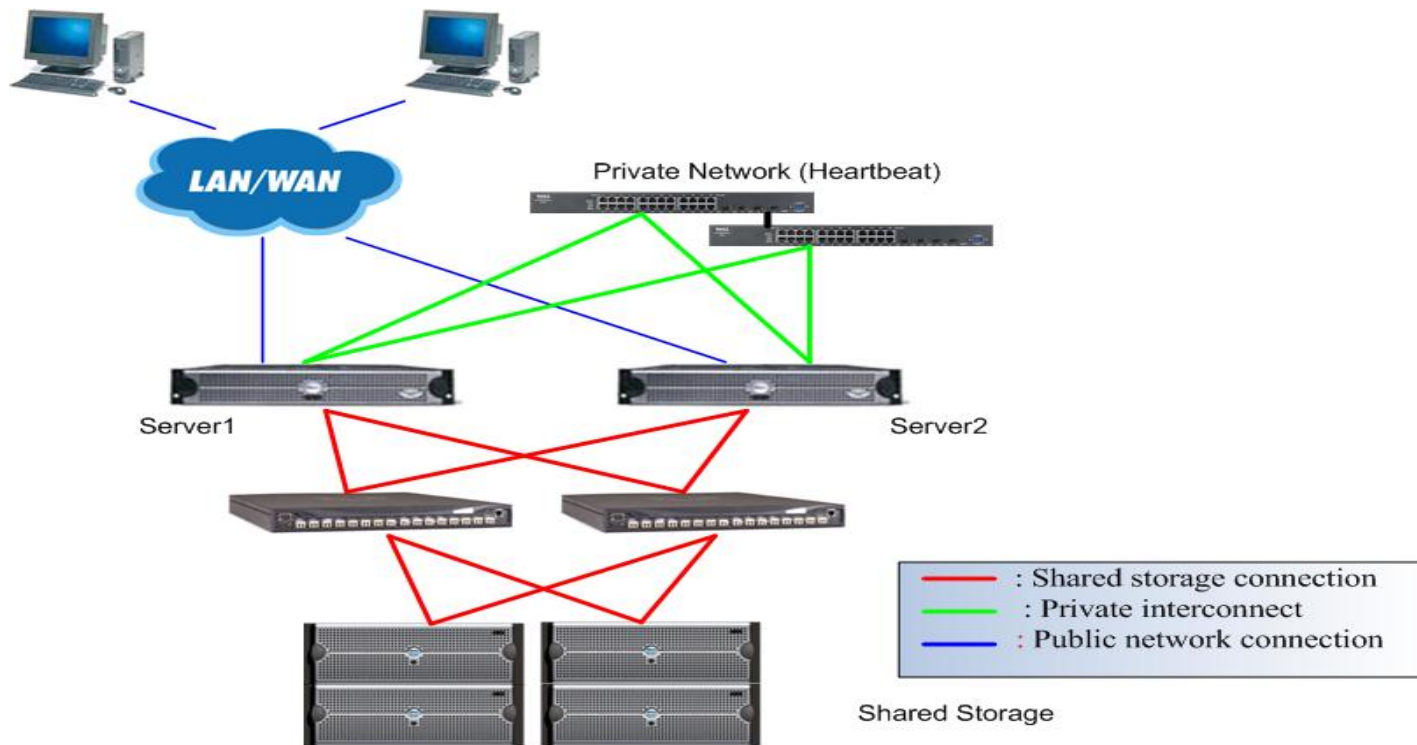
# Agenda

- Oracle 11g R2 Clusterware Architecture
- **Storage Configuration**
- Network Configuration
- Managing Oracle Clusterware
- Clusterware Troubleshooting
- RAC Database Case Study
- QA



# 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

---

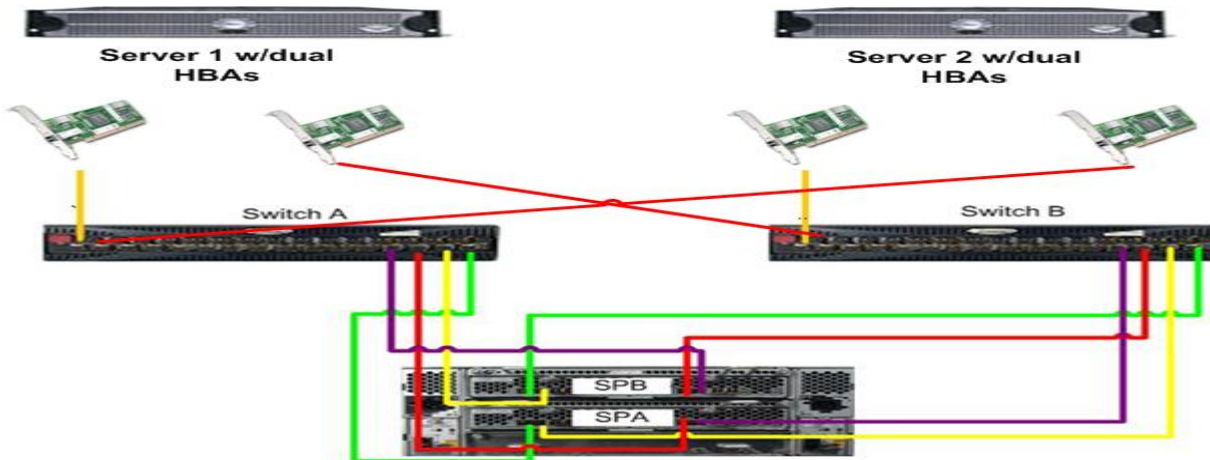
- Shared Storage Requirement:

- Shared storage for OCR and voting disk
- Types: block devices, RAW devices, OCFS/OCFS2, NFS on certified NAS (Oracle Storage Compatibility Program list)
- 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

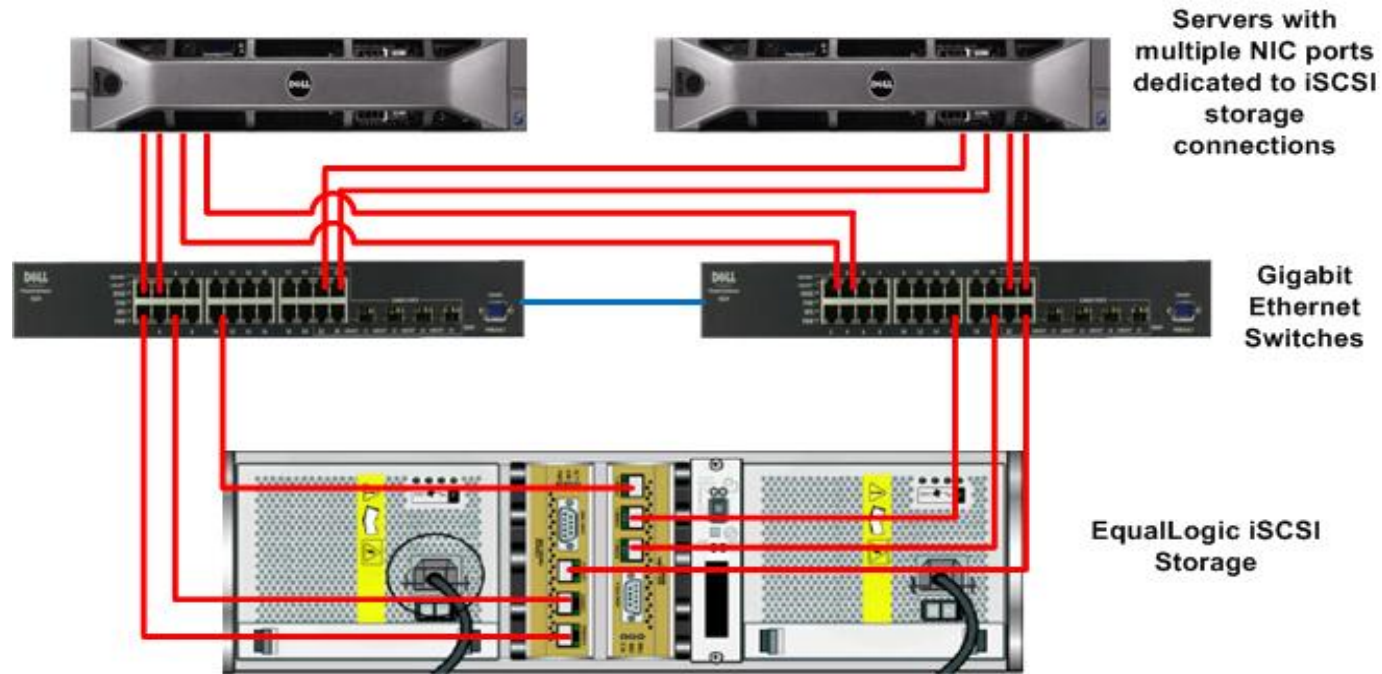


# Shared Storage Configuration

---

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-I/Os=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-I/Os  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 ASM Lib and Create ASM disks
  - . Consistent device naming and permission persistency in Linux
  - . Install the ASMLIB library: Load ASMLib rpms on Linux
  - . Configure ASM LIB: `service oracleasm configure`
  - . Create ASM disk: `service oracleasm createdisk DATA /dev/emcpowerd1`  
(`$/usr/sbin/asmtool -C -l /dev/oracleasm -n DATA -s /dev/emcpowerd1 -a force=yes`)  
`$/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

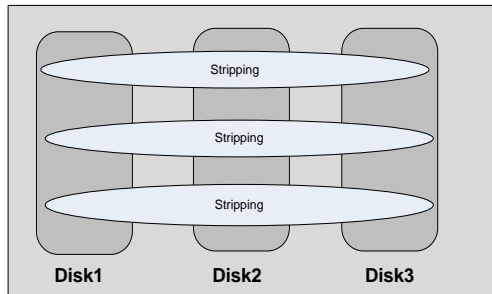
```
root@k4r815n1:/dev/oracleasm/disks
[root@k4r815n1 disks]# ls -l /dev/oracleasm/disks
total 0
brw-rw---- 1 grid asmadmin 8, 81 Nov 19 10:35 ACFS
brw-rw---- 1 grid asmadmin 8, 49 Nov 19 10:35 DATA
brw-rw---- 1 grid asmadmin 8, 129 Nov 19 10:35 DATA1
brw-rw---- 1 grid asmadmin 8, 65 Nov 19 10:35 FRA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



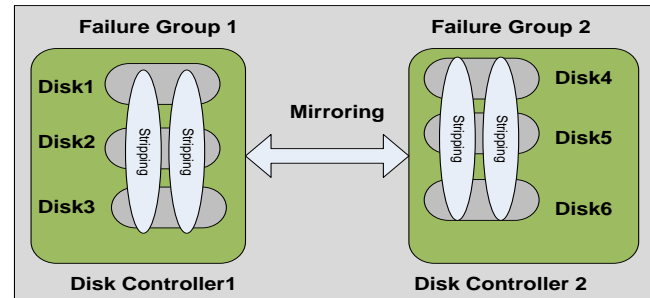
# Shared Storage Configuration

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

ASM Disk Group with External Redundancy

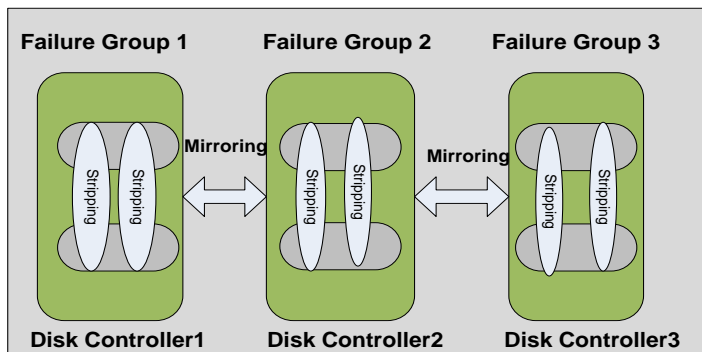


ASM Disk Group with Normal Redundancy



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

ASM Disk Group with High Redundancy



ASM failure groups

```
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 disks

High Redundancy (5 failure group): 5 voting disks

1 OCR + up to 5 copies: one per diskgroup

A quorum failure group: only used for OCR and votingdisk

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:

**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 uses 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:ACFS	PROVISIONED		20479	<input type="checkbox"/>
<input type="checkbox"/> ORCL:DATA	PROVISIONED		40959	<input type="checkbox"/>
<input type="checkbox"/> ORCL:FRA	PROVISIONED		20479	<input type="checkbox"/>

ASM Instances | **Disk Groups** | Volumes | ASM Cluster File Systems

You can choose to create a new disk group or add disks to an existing disk group. To create dynamic volumes, you need disk groups with 11.2 ASM compatibility.

Tip: To perform operations on a disk group, right mouse click on the row.

Disk Groups

Disk Group Name	Size (GB)	Free (GB)	Usable (GB)	Redundancy	State
ORAHOME	20.00	19.91	19.91	EXTERN	MOUNTED(2 of 2)
OCRVO			1.02	HIGH	MOUNTED(2 of 2)

- View Serviced Databases
- View Status Details
- Add Disks
- Edit Attributes
- Manage Templates
- Create ACFS for Database Home**
- Mount on Local Node
- Dismount on Local Node
- Drop
- Mount on All Nodes
- Dismount on All Nodes

**Create ACFS Hosted Database Home**

Create ACFS Hosted Database Home

Database Home Volume Name:

Database Home Mountpoint:

Database Home Size (GB):

Database Home Owner Name:

Database Home Owner Group:

```
[root@owirac1 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: DATA

Redundancy: Redundancy is achieved by storing multiple copies of the data on different failure groups. Normal redundancy is achieved from 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: FRA

Redundancy: Redundancy is achieved by storing multiple copies of the data on different failure groups. Normal redundancy is achieved from 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>alter diskgroup DATA_DG add disk
```

```
'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



# Agenda

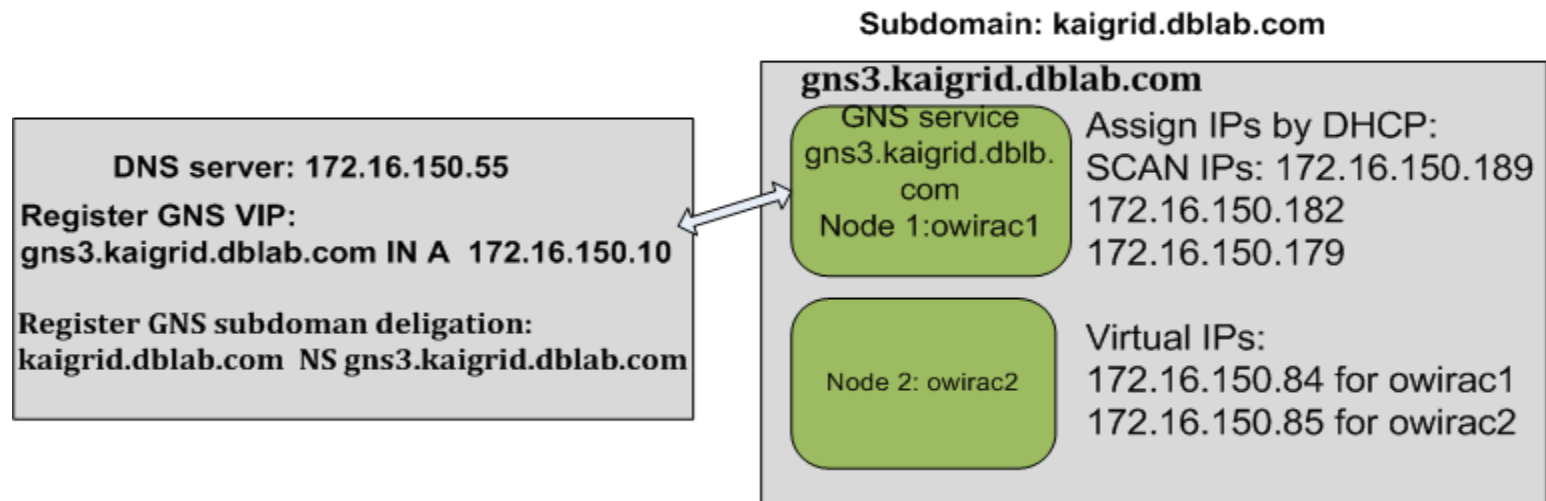
- Oracle 11g R2 Clusterware Architecture
- Storage Configuration
- **Network Configuration**
- Managing Oracle Clusterware
- Clusterware Troubleshooting
- RAC Database Case Study
- QA



# 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,
- Three ways to configure SCAN VIPs
  - DNS: add three entries to DNS:
  - GNS: assigned through DHCP of GNS(Grid Naming Services)
  - Added into /etc/hosts, only for test environment
- How GNS service works: an example: GNS assigns VIPs by DHCP  
xxxx.kaigrid.dblab.com → gns.kaigrid.dblab.com

```
scan_name IN 172.16.150.189
          IN 172.16.150.182
          IN 172.16.150.179
```



# Network Configuration

- How to setup GNS service through 11gR2 clusterware
  - On DNS server:
    - Configure GNS VIP in DNS : `gns3.kaigrd.dblab.com IN A 172.16.150.10`
    - Configure GNS Subdomain: `kaigrd.dblab.com NS gns3.kaigrd.dblab.com`
  - 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 kaigrd.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 kaigrd.dblab.com)

Configuring up Grid Infrastructure - Step 4 of 16

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:

# Network Configuration

---

## 11gR2 GI(Grid Infrastructure) installation

- Creates GNS service and GNS VIP
- Assigns VIPs and SCAN IPs through DHCP from 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
```

- Clusterware manages the high availability of GNS and GNS VIP

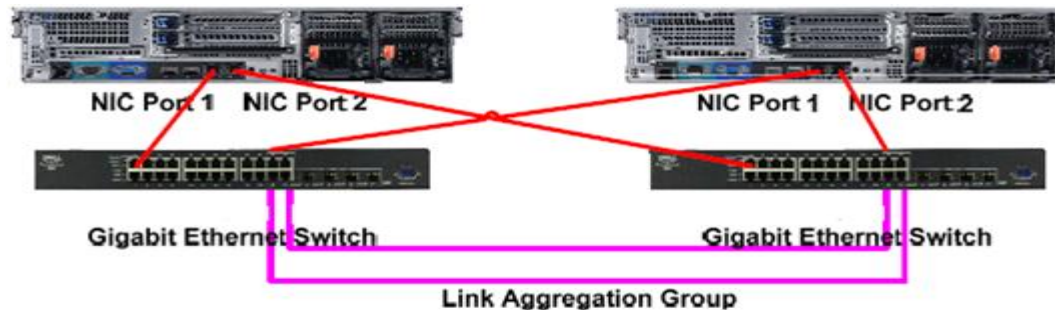


# Network Configuration

- Private Interconnection Configuration

- Fully Redundant Ethernet Interconnects:

- Two NIC cards, two non-roundtable interconnect switches



- NIC teaming to bond two network interfaces for failover

`ifcfg-eth1:`

`DEVICE=eth1`

`SLAVE=yes`

`ifcfg-eth2:`

`DEVICE=eth2`

`SLAVE=yes`

`ifcfg-bond0:`

`IPADDR=192.168.9.52`

`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



# Agenda

- Oracle 11g R2 Clusterware Architecture
- Storage Configuration
- Network Configuration
- **Managing Oracle Clusterware**
- Clusterware Troubleshooting
- RAC Database Case Study
- QA



# Managing Oracle clusterware

## ■ Managing voting disks

- Locate voting disk: `crsctl query votedisk css`
- Set odd number of voting disks
- Store Voting disks on ASM

One votingdisk per one failure group for redundancy

```
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             6bc4818a35d44ffebf95aaa31831cfb3 (ORCL:OCR1) [OCRVOTDSK]
Located 5 voting disk(s).
```

- Voting disk is backed up automatically in 11gR2
- crsctl commands to add, delete, replace voting disk.
- Restore voting disk: . Restore OCR if OCR corrupted
  - . Start crs in exclusive mode: `crsctl start crs -excl`
  - . Add new votingdisk: `crsctl replace votedisk +asm_disk_group`
  - . Restart crs



# Managing Oracle clusterware

## ■ Managing OCR

- Locate OCR:

```
grid@k4r815n1:~  
[grid@k4r815n1 ~]$ ocrcheck  
Status of Oracle Cluster Registry is as follows :  
      Version                :                3  
      Total space (kbytes)    :            262120  
      Used space (kbytes)     :             2860  
      Available space (kbytes):            259260  
      ID                      :       232227200  
      Device/File Name       : +OCRVOTDSK  
                          Device/File integrity check succeeded  
      Cluster registry integrity check succeeded
```

- Migrate OCR to ASM: a. add new ASM diskgroup  
b. *ocrconfig -add +new\_diskgroup*  
c. *ocrconfig -delete old\_location*
- Backup OCR: automatic: *ocrconfig -showbackup*  
manual: *ocrconfig -manualbackup*
- Restore OCR: stop clusterware: *crsctl stop*  
run restore command: *crs ocrconfig -restore*
- Diagnose OCR problem: **OCRDUMP** and **OCRCHECK**
- Export OCR: *ocrconfig -export*
- Import OCR: *ocrconfig -import*



# Agenda

- Oracle 11g R2 Clusterware Architecture
- Storage Configuration
- Network Configuration
- Managing Oracle Clusterware
- **Clusterware Troubleshooting**
- RAC Database Case Study
- QA



# 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 ocssd.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

- 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: [\\$CRS\\_HOME/log/<hostname>/cssd/ocssd.log](#) file in Node1:

```
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, seedhbimpd 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
[oahasd(14913)]CRS-8011:reboot advisory message from host: k4r815n2, component: mo163503, with ti
me stamp: L-2010-11-23-17:12:11.052
[oahasd(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



# Clusterware Troubleshooting

- 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 18:10:15.878
[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:

Jul 23 11:15:23 racdb7 logger: Oracle clsmom 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:
    - 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 step for 10g and 11gR1: `crsctl set css diagwait 13 -force`
    - 13 seconds is set by default in 11gR2
  - 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`



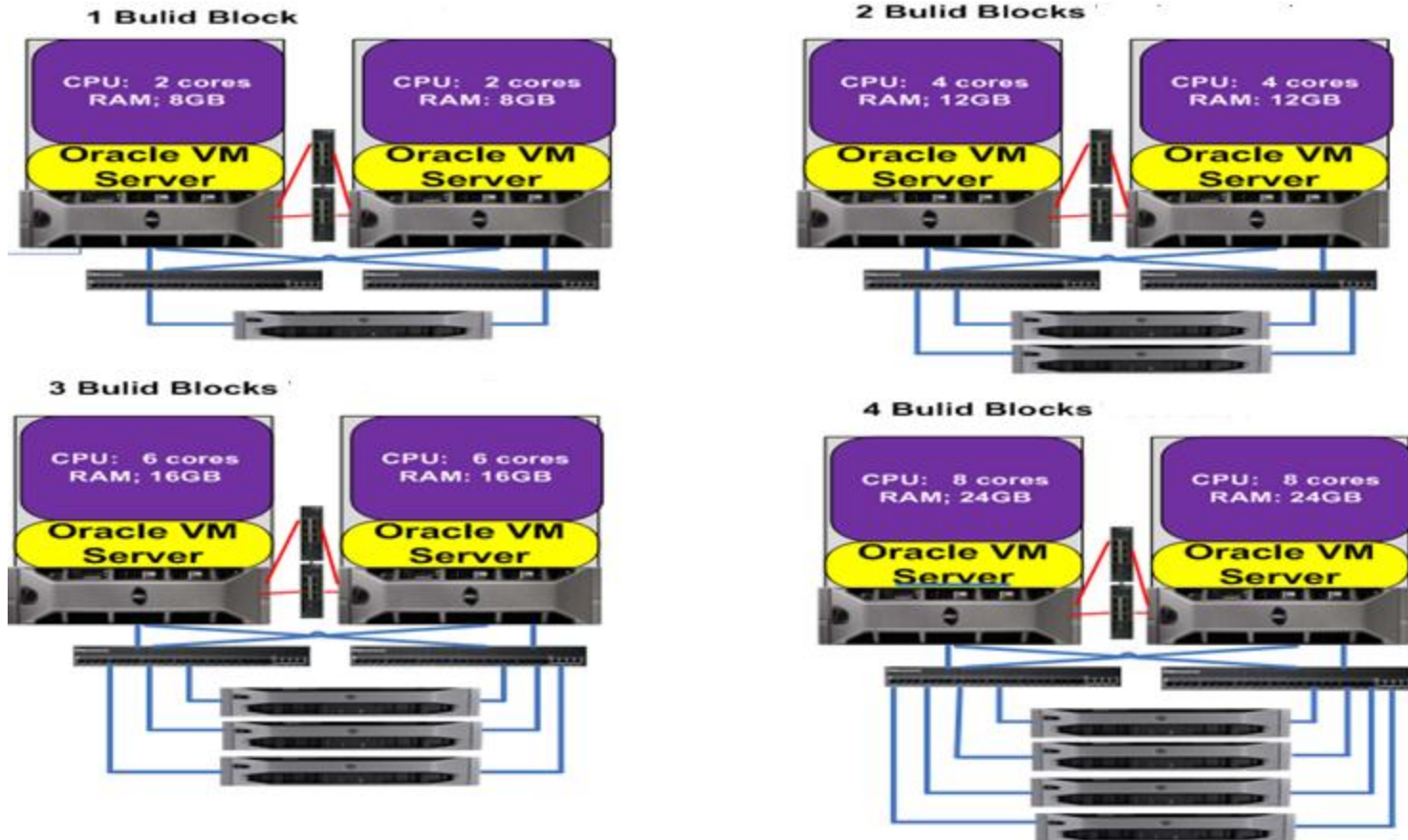
# Agenda

- Oracle 11g R2 Clusterware Architecture
- Storage Configuration
- Network Configuration
- Managing Oracle Clusterware
- Clusterware Troubleshooting
- **RAC Database Case Study**
- QA



# 11gR2 RAC Case Study

## Scalable Data Warehouse Building Blocks Solutions



# 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/>

## Oracle ACE Director Kai Yu's Oracle Blog

Share Oracle Tips and Tricks and Related News



[About Me](#) [Presentations](#) [Publications](#) [Resources](#)

Received the 2011 OAUG Innovator of the Year Award Posted by: [kyuoracleblog](#) | April 22, 2011

I AM A MEMBER OF



ARCHIVES

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





The **Oracle Applications User Group (OAUG) Innovator of the Year award** recognizes a member company that has demonstrated an innovative technique using the standard functionality of an Oracle Application. On April 11, 2011, at the Collaborate 11 conference, OAUG announced this year's innovator award winner Dell Inc, for its pioneer breakthrough method for standardizing the Oracle Applications deployment in the virtual environment.

Oracle VM provides the server virtualization that not only enables high availability and scalability, also simplifies and standardizes the deployment for Oracle Applications E-Business Suite. To leverage Oracle VM, the Oracle Applications E-Business Suite need to be virtualized and a new deployment method needs to be established on the virtualized environment. Although Oracle provides the certification of running Oracle Applications on Oracle VM and Oracle Applications vision instance virtual machine template to build the vision demo Oracle Applications system on Oracle VM, there is no documented method to