**Shared Nothing Architecture:**

Shared-nothing clusters are where storage is not shared and data must be either replicated or segmented across the cluster.

Used for workloads that can be easily and predictably divided into small units that can be spread across the cluster in parallel.

**Shared Everything Architecture:**

These clusters can perform these tasks but also offer increased flexibility for varying workloads.

Oracle provides load balancing and failover capabilities by using this.

All nodes must share the same processor architecture and run the same operating system.

**Real Application Clusters**

Uses a Shared everything architecture.

Each instance has its own redo log thread & undo instance.

The users/application view the system as a single entity.

Services define how the users/apps are going to be connected.

Cache Fusion enables merging buffer cache for all instances to make it appear as one unit.

Data blocks can be transferred directly from one instance buffer cache to another instance buffer cache.

Every instance has its own set of background processes.

**What is Clusterware??**

* Oracle Clusterware enables servers to communicate with each other, so that they appear to function as a collective unit.
* This combination of servers is commonly known as a cluster.
* Although the servers are standalone servers, each server has additional processes that communicate with other servers.
* In this way the separate servers appear as if they are one system to applications and end users.

**Clusterware Versions**

|  |  |
| --- | --- |
| **Oracle Cluster Manager**  **(Supports Linux & Windows)** | **Version** |
| Oracle Cluster Manager  (Supports Linux & Windows) | 9i |
| Oracle CRS  (Supports all OS. Made mandatory for building 10g RAC on Linux) | 10.1 – CRS Home  - ASM Home  - DB Home |
| Oracle Clusterware  (Renamed CRS to Clusterware. Supports HA Operation & Portable) | 10.2 – CRS Home  - DB + ASM Home |
| Grid Infrastructure  (Combination of Clusterware software + ASM Binaries) | 11.2 – Grid Home (CRS+ASM)  - DB Home |

**Shared Storage Types**

* NAS – Network Attached Storage
* SAN – Storage Area Network
* ISCSI – Internet Small Computer System Interface

**Software Requirements for 11gR2**

* Same version & same Architecture of OS on all nodes.
* Same version of Clusterware software on all nodes.
* Same or lower version of RDBMS software on all nodes compared to clusterware software.

**RAC Components**

* Cluster Software
* Private network OR Interconnect & Public Network
* Shared Storage

Clusterware software:

It coordinates and manages each node membership within the cluster by treating all nodes as a single large logical server.

Shared Storage:

It is a dedicated type of external storage system which is simultaneously accessible by multiple nodes of a cluster.

Public Network:

For communicating with the shared storage.

**Private Interconnect**

Prt?

Private Interconnect:

It is a high bandwidth and low latency inter-**communication setup used for transferring the heartbeat messages among all the nodes of the cluster**

* Identifying health, status and message synchronization among all the instances.
* Used for transferring global resource lock requests among all the instances.
* Heavily used for transferring data blocks from one instances buffer cache to other instance buffer cache.
* The more instance’s communicate, more traffic on the network. Hence, it is recommended to have high bandwidth network as well as redundant interconnect.

**Clusterware software:**

**Software that provides various interfaces and services for a cluster. Typically, this includes capabilities that:**

* + Allow the cluster to be managed as a whole
  + Protect the integrity of the cluster
  + Maintain a registry of resources across the cluster
  + Deal with changes to the cluster
  + Provide a common view of resources

**Clusterware software:**

A key part of Oracle Grid Infrastructure.

Integrated with Oracle Automatic Storage Management (ASM).

The basis for ASM Cluster   
File System (ACFS)

A foundation for Oracle   
Real Application Clusters   
(RAC).

A generalized cluster   
infrastructure for all kinds   
of applications.

**Oracle Clusterware Architecture and Services**

**Shared disk cluster architecture supporting application load balancing and failover.**

**Services include:**

* Cluster management
* Node monitoring
* Event services
* Time synchronization
* Network management
* High availability

**Goals for Oracle Clusterware**

* Easier installation
* Easier management
* Continuing tight integration

with Oracle RAC

* ASM enhancements with

benefits for all applications

* No additional clusterware

required

**Oracle Clusterware Networking**

* + Each node must have at least two network adapters.
  + Each public network adapter must support TCP/IP.
  + The interconnect adapter must support:

All platforms use Grid Interprocess Communication (GIPc)

**Interconnect Link Aggregation: Multiswitch**

* + Redundant switches connected with an Inter-Switch Trunk may be used for an enhanced highly available design.
  + This is the best practice configuration for the interconnect**.**

**2 Node RAC Setup**

* Public/Private network – Under control of O/S.
* Virtual network – Under control of Clusterware software.
* Public network communicates with Shared storage.
* Public & Virtual network – Same subnet mask.

**RAC Processes**

**Architecture Components**

**Prt?**

**Private Interconnect**

Used for internode communication and also identifying the health status of each node and transferring the global resource request locks, etc.

**PN?**

**Pub?**

**Public Network**

Used for client/server communication & also for accessing the shared storage.

Both public & private network are under the control of node-specific OS.

**GRD?**

**Global Resource Directory (GRD)**

A new memory component which is a part of shared pool and can be seen only in RAC specific instances.

* + Oracle automatically maintains GRD consistency among all the instances of a common DB by using LMON background process.
  + GRD maintains metadata info of data blocks which is present in DBC like:

- SCN#

- Data Block Address.

- Location of the most recent version of the block.

- Mode of the Block (NULL, SHARED, EXCLUSIVE)

- Role of the block (LOCAL, GLOBAL)

- Type of Block Image (CURRENT, CONSISTENT READ, PAST IMAGE)

* GRD is maintained by communicating with GES & GCS.

GES?

* **Global Enqueue Services (GES):**

Coordinates with the Global Lock Requests and non-cache fusion operations with the help of LMD & LCK background processes.

**GCS?**

**Global Cache Services (GCS)**

Maintains by coordinating with cache fusion operations with the help of LMSn background process.

**CF?**

**Cache Fusion**

Cache Fusion is the remote memory mapping of Oracle buffers

* is the process of transferring the available data blocks from 1 instance buffer cache to other instance buffer cache by using existing private network.
* Avoids expensive hardware intense disk reads
* Has 2 phases:
  + Phase I was supported in Oracle 8i OPS for data read. Disk I/O had to be performed for DML operations.
  + Phase II supported both disk read & DML operations.
* Cache Fusion is also called 'Soft Ping'.

**RAC Background Processes**

**LMS?**

**LMSn (Global Cache Service Process):**

Actual process of transferring data blocks from one instance buffer cache to another instance buffer cache.

Controlled by gcs\_server\_processes parameter

LMON?

**LMON (Global Enqeueue Service Monitor):**

Responsible for maintaining GRD consistency among all the instances of a database.

LMD?

**LMD (Global Enqueue Service Daemon):**

Responsible for maintaining global lock requests.

LCK?

**LCK (Lock Process):**

Handles all the non-Cache Fusion operations

DIAG?

**DIAG:**

Updates the diagnostic info into trace files and alert log files whenever server process or any other background process needs diagnosability.

**11gR2 New Features**

NF?

• New (complex) network requirements

• 2 Oracle Homes (instead of 3)

– ASM and CRS together in GRID home

– Must be different ORACLE\_BASE

• OCR/Vote not supported on raw (Backward compatible)

• Use ASM or CFS instead of RAW.

• Cluster Time Synchronization Service

• Software only Grid install possible.

• RAC OneNode.

* Oracle Restart.
* Instance Caging
* Upto 30 Standby databases
* Active Data Guard.
* SCAN
* RMAN - Automatic Block Recovery.
* Grid Plug & Play (GPNP)
* New File System – ACFS
* ACFS Dynamic Volume Manager (ADVM)
* Database Replay.

**Instance Caging**

Cag?

Instance Caging limits the amount of CPU an Oracle database instance consumes, using the Oracle Database Resource Manager and the cpu\_count parameter:

* The Resource Manager limits the amount of CPU that the database instance consumes.
* The cpu\_count parameter specifies the limit.
* Provides CPU resource allocation control to a DBA.
* Rather than using logical/physical partitioning of a large box resorting to usage of tools/technologies such as VMware, AIX LPAR or Sun containers etc.

**Active Dataguard**

AD?

Active Dataguard is a new feature of 11gr2 which eliminates the DR to remain idle during Production workloads.

An active standby can offload ad-hoc queries, reporting, and fast incremental backups from the primary database, improving performance and scalability while preventing data loss or downtime due to data

* corruptions database
* and site failures,
* human error, or
* natural disaster.
* Protection
* Availability
* Return On Investment
* Performance

**Rac OneNode**

Onenode?

Rac OneNode is a single instance of Oracle RAC running on one node in a cluster.

Has all the features of RAC databases.

Good Scalability and High Availability options as compared to Single Instance databases because additional nodes can be added if the existing single node becomes overloaded.

In 11.2, Oracle RAC One Node will be available on Linux

only.

**Oracle Restart**

Excellent High Availability solution for:

-- Database Instance -- ASM

-- Listener -- Services

Restart?

* Automatically restarts after a hardware or software failure or whenever your database host restarts.
* Runs out of the Grid Infrastructure home which is separate from the database home.
* Server Control (SRVCTL) is the command line interface to manage Oracle processes that are managed by Oracle Restart on a standalone server.

**RMAN - Automatic Block Media Recovery**

* Automatic Block Media Recovery provides fast and transparent block recovery functionality in a real-time query physical standby database.
* Corruption on either Primary or Standby is handled implicitly.
* Blocks are automatically copied from the alternative database to restore the corrupt data block.
* An end user will probably not even notice the Block Corruption on the Primary Database.

E.g.:

Reading datafile '/home/oracle/physt/users01.dbf' for corruption at rdba: 0x01000083 (file 4, block 131)

Reread (file 4, block 131) found same corrupt data (no logical check)

Automatic block media recovery requested for (file# 4, block# 131)

**Database Replay**

Replay?

* Recreate actual production database workload in test environment
* Identify, analyze and fix potential instabilities before making changes to production
* Capture Workload in Production
  + Capture full production workload with real load & concurrency info
  + Move the captured workload to test system
* Replay Workload in Test
  + Make the desired changes in test system
  + Replay workload with production load & concurrency
* Analyze & Report
  + Errors
  + Data divergence
  + Performance divergence

**Database Replay: Supported Changes**

**Grid Plug and Play**

* + In previous releases, adding or removing servers in a cluster required extensive manual preparation.
  + In Oracle Database 11*g* Release 2, GPnP allows each node to perform the following tasks dynamically:
    - Negotiating appropriate network identities for itself
    - Acquiring additional information from a configuration profile
    - Configuring or reconfiguring itself using profile data, making host names and addresses resolvable on the network
  + To add a node, simply connect the server to the cluster and allow the cluster to configure the node.

**SCAN**

Scan?

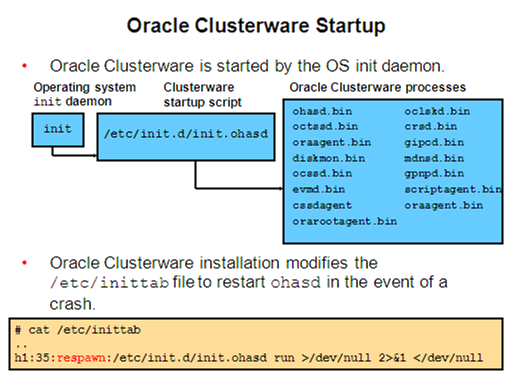
* + Single Access Client Name allows client connections to connect to the database irrespective of the number of nodes in the cluster.
  + No change required in the connection strings even when adding or deleting nodes.
  + SCAN IP’s must be on the subnet as that of public IP address.
  + Provides Intelligent Load Balancing features for client connection.
  + 3 SCAN Listeners for the entire cluster.

**ACFS & ADVM**

ACFS?

* + New shared file system (ASM Clustered File System) used for storing non-database, general purpose files in the ASM Diskgroups.
  + ASM Dynamic Volume Manager helps manage the FS.
  + Leverages ASM technology for volume management.
  + Integrated with Oracle Clusterware for cluster support.
  + Managed by:
    - Native OS commands
    - acfsutil CLI
    - ASMCA
    - OEM

**Clusterware Startup Sequence**



**Startup Sequence – In Depth**

Init?

**Startup Sequence:**

The init process invokes OHAS: starts $GRID\_HOME/bin/ohasd.bin

Logs in the activities in $GRID\_HOME/log/<node>/ohasd/ohasd.log

**OHASD >>** oraagent.bin (started as Grid s/w owner)

cssdmonitor (started as Root)

cssdagent (started as Root)

orarootagent.bin (started as Root)

All four are located in $GRID\_HOME/bin

orarootagent >> crsd.bin (started as Root)

diskmon.bin (started as Grid s/w owner)

octssd.bin (started as Root)

oraagent >> evmd.bin (started as Grid s/w owner)

evmdlogger.bin “

gipcd.bin “

gpnpd.bin “

mdnsd.bin “

cssdagent starts ocssd.bin

**CRSD >>** oraagent.bin (started as Grid s/w owner)

oraagent.bin (started as oracle)

orarootagent.bin (started as Root)

Log of all three are located in $GRID\_HOME/log/<node>/agent/crsd/

oraagent (Grid owned) >> Clustered ASM Instance

ONS

SCAN Listener

Node Listener

oraagent (Oracle owned) >> Databases

Services

orarootagent >> network resources

SCAN VIP

Node VIP

GNS VIP if configured

GNS Daemon

ACFS Registry

**Cluster Architecture Processes/Files**

OHASD (Oracle High Availability Services Daemon):

First & the only daemon which is started by the parent init process & in turn it is responsible for starting many agents and daemons. Reads OLR file.

Olr?

OLR (Oracle Local Registry):

Contains startup sequence of all the child daemons

Crsd?

CRSD (Cluster Ready Services Daemon):

Responsible for performing cluster configuration & Supports HA by reading OCR file.

Contains complete cluster information

Cssd?

CSSD (Cluster synchronized Services Daemon):

Responsible for updating the node membership into VD.

VD (Voting Disk):

Maintains the node membership. Contains their IP & subnet masks.

CSSD sends heartbeat messages every second & receives the same. Stores into VD.

EVMD (Event Manager Daemon):

Responsible for publish & subscribe event notifications to other nodes which are raised by CRSD.

OCTSSD (Oracle Cluster Time SSD):

Maintains time consistency on all nodes or instances.

GPNPD (Grid Plug & Play Daemon):

Responsible for updating GPNP profile whenever basic cluster configuration is changed.

GPNP Profile:

Maintains basic cluster information like cluster node names, their corresponding subnet masks, ASM Diskstring, location of VD file, location of ASM spfile.

GSD (Global Service Directory):

Performs administrative tasks whenever an GUI application like DBCA or NETCA is invoked.

VIP (Virtual IP):

Registered as a service into OCR & maintains the status also.

Till 11.1, we used node-specific private IP, public IP & VIP

From 11.2, we require 2 more IP's on the same subnet of public IP used for SCAN

**Cluster Administration**

Cluster Level files: OCR & VD

Node Level Files: OLR & GPNP Profile

Oraagents run with 'Oracle' user privileges & orarootagents run with 'root' privileges.

Lower or HAS stack needs access to both OLR & GPNP profile and CRS/upper stack needs access to OCR & VD files.

Every node specific daemon will connect to its peer daemon on other nodes.

**11gR2 New Features: Cluster**

NF?

ASM must start *before CRS*

• More components to manage

• Easier to start & stop nodes or entire cluster

su - grid

su

crsctl stop cluster –all

**11gR2 New Features: Cluster**

**crs\_stat (my favorite) is deprecated**

– No longer shows instance status

• Replace by **crsctl status resource –t**

– Run as grid

– Many more components

• Disk groups, network, scan, etc

– Local vs. Cluster resources

**Pre-Installation Tasks – root**

**Pre?**

* Install same OS on all nodes – login as root
* Configure VIP (eth0:0) on top of Public IP using ‘neat’
* Update /etc/hosts file with all node related Private IP mapping to Private Hostnames
* Configure DNS Server & DNS Clients
* Configure Shared Storage
* Create groups, users, directory structures & assign permissions
* Install required OS/Clusterware packages
* Update kernel parameters & sync date/time on all nodes
* Update .bash\_profile with the required env variables
* Configure SSH for password-less connectivity
* Perform pre-cluster verification check
  + Component level
  + Cluster level

**Installing Grid Infrastructure**

Grid?

* Start **runInstaller** for installing Grid Infrastructure
* Verify the Cluster after completion with **crsctl**
* The Clusterware & ASM Instance are now configured.

**RAC related Parameters**

Para?

* Identical Parameters:

These parameters must have the same value for all the instances of a common database.

Eg: cluster\_database db\_name db\_files

control\_files db\_recovery\_file\_dest

* Non-identical Parameters:

These parameters have different values for each and every instance.

Eg: thread instance\_number

instance\_name undo\_tablespace

**RDBMS Software Install on RAC**

* Edit the .bash\_profile to set the environment variables to point to the RDBMS Home.
* Start the runInstaller from 11g database setup from one of the nodes
* The files are copied to the other nodes during install.
* Create diskgroups for the database & mount them on both nodes.
* Create required directories for storing data in the diskgroups with either sqlplus or ASMCMD

**Database Creation on RAC**

* On any of the node, create temporary parameter file.
* Connect to the sqlplus & create global spfile using temp file
* Create node-specific pfile pointing to the spfile
* Start the instance & create database in shared storage
* Create 2nd node specific Redo log thread & Undo tablespace
* Modify cluster specific parameters. Set them to TRUE
* On node 2, create pfile pointing to global spfile & start instance
* Run database post-installation scripts

**Registering resources into OCR**

* **SRVCTL Utility:**

It is used to manage database/cluster resources. Performs starting/stopping, enabling on boot, etc.

srvctl add database –d <db\_name> -o <oracle\_home>

-a <asm\_diskgroups>

srvctl add instance –d <db\_name> i <instance\_name> -n <node\_name>

**Clusterware Files**

Cluster Related Files:

* Oracle Local Registry (OLR)
* Oracle Cluster Registry (OCR)
* Voting Disk (VD)

GPNP Profile

**OLR (Oracle Local Registry)**

Oracle automatically performs OLR backup at the time of execution of root.sh script and stores the info in the location:

$GRID\_HOME/<node\_name>/backup\_date\_time.olr

Actual File Location:- $GRID\_HOME/cdata/<node>.olr

Manual Backup:- ./ocrconfig -local -manualbackup

Check Integrity:- ./ocrcheck –local

Change Backup Location: ./ocrconfig –local –backuploc /home/oracle

‘Local’ keyword point to OLR

Restore OLR Backup:- ./ocrconfig –local –restore <backup\_file\_location>

**OCR (Oracle Cluster Registry)**

**OCR?**

* It is a critical shared file which completes cluster info like cluster node names, corresponding IP addresses, CSS Parameters, OCR Backup info and registered resources like node apps, ASM Instances, Databases, Database Instances & DB Services, etc.
* Every node specific CRSD daemon is responsible for maintaining and updating OCR file.
* CRSD daemon automatically brings up all the services which have got registered into OCR as it gets up

To check Integrity:

./ocrcheck

Logfile Location:

$GRID\_HOME/log/cdata/crsd

Backup Location:

$GRID\_HOME/cdata/<cluster\_name>

OCR Backup Methods:

1) Automatic: Oracle performs OCR backups every 4 hours of interval & stores info into master node.

Oracle retains latest three 4 hour interval backups, 1 day backup & 1 week backup & purges remaining ones.

Change backup location:

./ocrconfig -backuploc <new\_location>

Restore OCR:

./ocrconfig -restore <backup\_loc>

Physical Backup of OCR:

cp /ocfs/ocr /home/oracle/ocr.bkp

dd if=/dev/raw/raw5 of=home/oracle/ocr.bkp bs=100MB

Logical Backup of OCR:

./ocrconfig -export <backup\_loc>

./ocrconfig -import <backup\_loc>

**OCR (Oracle Cluster Registry)**

Take Backups whenever:

Adding or deleting a node

Registering/unregistering a node into OCR

Cluster configuration get modified

Registering Services into OCR:

SRVCTL [recommended]

DBCA [Till 10.2]

OEM

OCR Multiplexing:

Avoids OCR loss & complete cluster downtime due to a single point of failure.

In 10.2 - max 2 locations

From 11.2 - max of 5 locations [1-primary & 4 mirror]

Also supports storage of OCR in ASM Diskgroups, the redundancy of the file will depend on redundancy level of the diskgroup.

**VD (Voting Disk)**

Vd?

Another crucial and shared file which contains all the cluster node membership info & their IP's and subnet masks, etc.

Every node specific CSSD daemon is responsible for sending heartbeat messages to all the other nodes for every one second and receive the heartbeat messages from other nodes and stores the response into VD.

Voting Disk Backup:

cp /ocfs/vd

dd if=/dev/raw/raw5 of=home/oracle/vd.bkp bs=100MB

Voting Disk Restore:

./crsctl -restore -votedisk <backup\_loc>

From 11.2, Oracle supports storing VD in ASM Diskgroup.

Oracle recommends atleast 1 failgroup for external redundancy, 3 for normal redundancy & 5 for high redundancy.

17. How do I identify the voting disk location?  
Ans:  
# [crsctl query css](http://satya-racdba.blogspot.com/2010/01/crsctl-commands.html) votedisk

**Voting Disk Multiplexing**

To avoid VD loss, complete cluster downtime due to Single Point of Failure, Oracle supports multiplexing of VD.

10g – 31 locations

11.2 – 15 locations

From 11.2, Oracle automatically maintains VD backup into OCR file itself.

For locating all the VD allocated:

$GRID\_HOME/bin/crsctl query css votedisk

**GPNP Profile**

* It contains basic cluster information like VD locations, ASM spfile location, ASM disk string, all nodes subnet mask information, etc.
* Every node specific GPNP daemon is responsible for updating GPNP Profile whenever the basic cluster info is changed & distribute the same to others.

Backup Location:

$GRID\_HOME/cdata<node\_name>/peer/gpnp/profile.xml

Actual Location:

$GRID\_HOME/cdata/peer/gpnp/profile.xml

**Cluster Management Utils**

* SRVCTL
* CRSCTL
* OCRCHECK
* OCRDUMP
* OIFCFG
* OLSNODES
* CLUVFY

**SRVCTL**

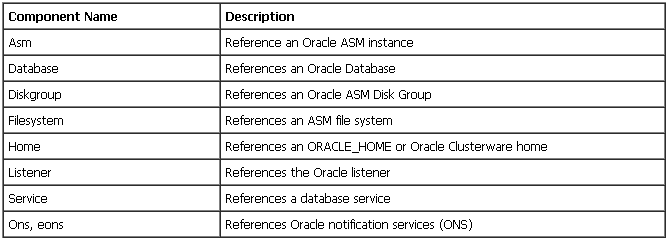
It is used to manage database/cluster resources. Performs starting/stopping, enabling on boot, etc.

srvctl add database –d <db\_name> -o <oracle\_home>

-a <asm\_diskgroups>

srvctl add instance –d <db\_name> i <instance\_name> -n <node\_name>

The following table lists the components that can be managed by the **srvctl** command



**CRSCTL**

* The **crsctl** utility is used, along with the **srvctl**, to manage and monitor Oracle Clusterware resources and components.
* **crsctl** is stored under the GRID\_HOME/bin directory.

Common crsctl commands:

$ crsctl check crs  
$ crsctl stat res –t

$ crsctl start/stop cluster [upper stack]

$ crsctl start/stop has [lower stack]

$ crsctl start/stop crs [Both lower & upper stack]

$ crsctl disable/enable crs

$ crsctl query css votedisk [Locates all Voting Disks]

crs\_\* commands are deprecated in 11.2

**Cluster Management Utils**

**OCRCHECK**

The **ocrcheck** utility is used to validate the integrity of the OCR and the OLR. Use the **ocrcheck** command whenever you have made any changes to your cluster (such as changing the VIP for example).

Arguments for OLR & OCR Integrity Check:

**-local** – Check the OLR integrity

**-config** – Display configured OCR’s.

**OCRDUMP**

The ocrdump utility allows you to dump the contents of the OCR or the OLR to a file or to stdout.

You can then read the resulting output for diagnostic and administration purposes.

ocrdump /tmp/outputfile.ocr

**OCRCONFIG**

The**ocrconfig** command is used to manage the OCR.

**ocrconfig** provides the ability to import, export, add, delete, restore, overwrite, backup, repair, replace, move, upgrade, or downgrade the OCR.

Arguments for OLR & OCR changes:

**-local** – Check the OLR integrity

**-config** – Display configured OCR’s.

**OIFCFG**

The Oracle Interface Configuration Tool (**oifcfg**) is used define and administer network interfaces such as the public and private interfaces.

Allocate network interfaces to components

Deallocate component network interfaces

Direct components to use specific network interfaces

Display component network configuration information

**OLSNODES**

The **olsnodes** command provides you information about the nodes on the cluster such as:

Node Names VIP’s assigned

Node Status Node Type

**CLUVFY**

The CVU is used to check the cluster at various times such as installation, during patching, adding or removing nodes and other types of cluster related change operations.

CVU will check the integrity of the cluster

cluvfy comp ssa –n <nodes> [Shared Storage Check]

cluvfy comp nodecon –n <nodes> [Node Connectivity Check]

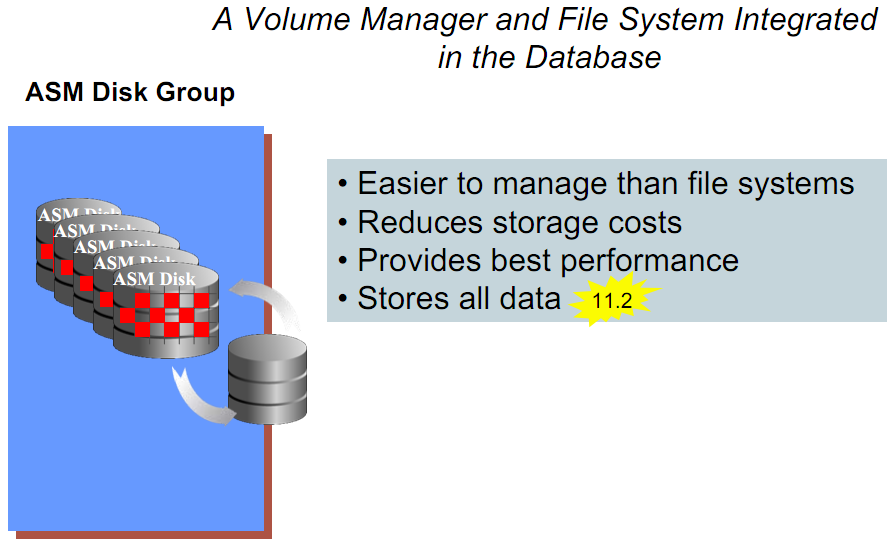
cluvfy comp asm [ASM Integrity]

cluvfy comp scan [Check SCAN config]

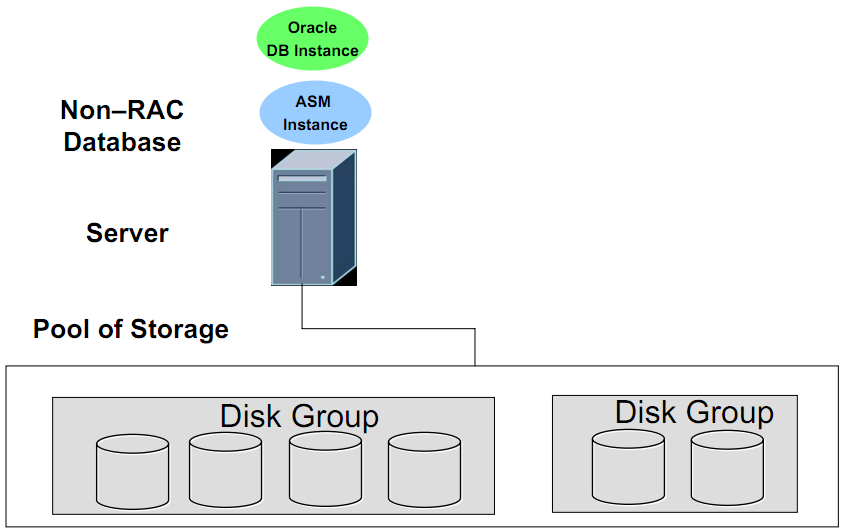
cluvfy comp clocksync [checks Clock Synchronization]

**ASM Overview**

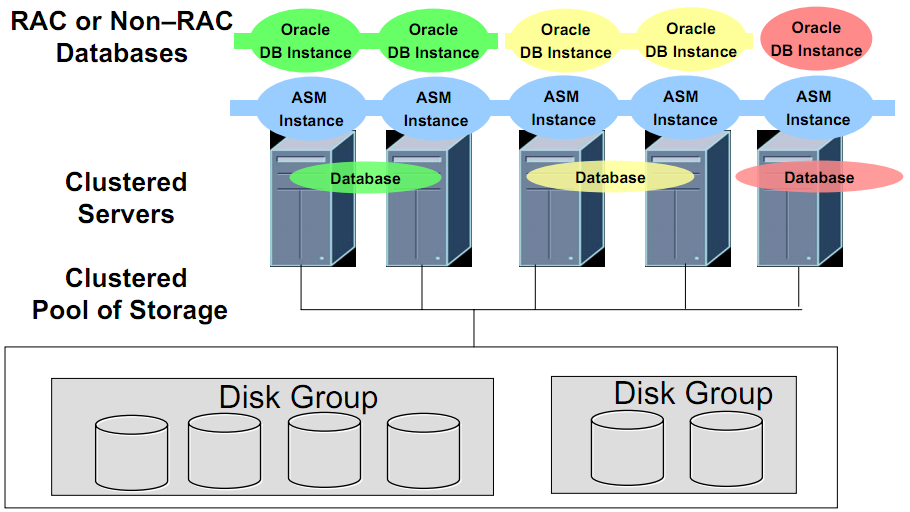
**Automatic Storage Management**



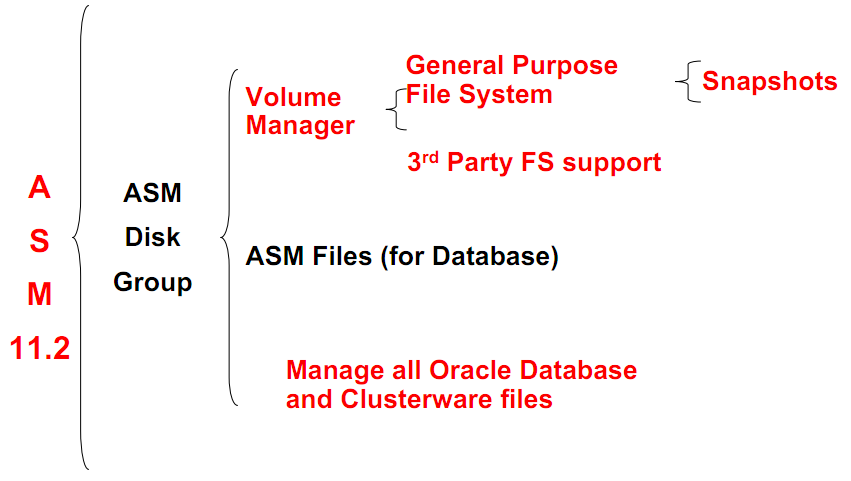
**ASM Process Architecture**



**ASM Cluster**



**Extending ASM to support all files**



**ASM 11gR2 New Features**

**ASM Dynamic Volume Manager**

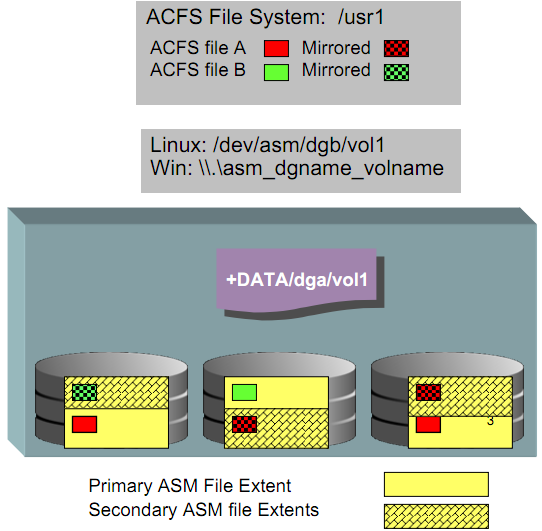
**Creating a Dynamic Volume**

**Oracle Clustered File System 2 (OCFS2)**

**ASM Clustered File System**

**Managing ACFS**

**Creating ACFS**

**Leveraging ASM Mirroring and Striping** 

**ACFS Read-Only Snapshots**

• Dynamic, fast, space efficient, “point in time” copies of ASM file system files

• An enabler for:

• On-line backups

• On-line, disk-based, file backup model using snapshots and individual file recoveries

• Up to 63 snapshot images per ASM file system

• Policy based snapshots:

• Schedule snapshots on an interval basis: every 5 seconds, every 30 minutes, daily, … with recycling (using EM)

• ACFS CLIs support creation and removal of snapshots

• ACFS Snapshot functions integrated with EM

**RMAN Overview**

Unlike physical backups taken in ‘backup mode’ which produces excessive redo logs & hampers performance till the backup finishes, RMAN poses no such implications.

Knows Oracle blocks

RMAN Components:

Target Database

Repository (Control File or Catalog)

RMAN commands are platform independent.

A single catalog can register multiple databases.

Backs up:

Control files Archivelog files

Database files Spfile

Won’t backup:

Redo log files listener.ora

Pfile tnsnames.ora

If target is on ASM, then RMAN is the only backup option.

Provides Block Media Recovery: Avoids ORA-1578 (Block Corruption)

Introduced a process – CTWR (Change Tracking Writer Process)

> recover datafile <file\_id> block <block\_id>;

> alter database enable block change tracking using file <file\_name>;

**Configuring Archiving/Flashback**

Flashback pre-requisites:

* Archivelog mode is mandatory.
* Flash Recovery Area Enabled
* For RAC, flash recovery area must be on a Clustered File System or ASM.

Steps:

* Configure flash recovery area by setting db\_recovery\_file\_dest\_size and db\_recovery\_file\_dest.
* Next step is to enable archivelog mode and then to turn on flashback.

To perform this, shutdown the database & start 1 instance in the mount mode to enable archiving & flashback.

* Open the instance & verify using:

SELECT LOG\_MODE,FLASHBACK\_ON FROM V$DATABASE;

Startup remaining instances.

**Instance Recovery in RAC**

* All instances have their own redo log thread & thus have their own set of archive logs.
* Every instance should have access to the redo logs + archivelogs of the other instances so as to recover it in case of a crash.

2-step recovery process in RAC:

* SMON of recovering instance performs roll forward & rollback for recovery of failed instance.
* LMON of recovering instance performs GRD recovery of failed instance.

**Virtual IP**

**Vip?**

* A dummy network created on top of the public network & recommended for the user session to connect with RAC database.
* Instead of users waiting for TCP/IP timeout which has a duration of 15mins, the corresponding VIP failsover immediately to the other surviving node by CRSD.
* Till 11.1, we required each node specific VIP which is on the same subnet mask as the Public IP.
* From 11.2, we require 3 more unused VIP’s on the same subnet of Public IP for SCAN VIP’s & all are resolvable to a single SCAN name.

**SCAN – Single Client Access Naming**

• A single *hostname to access the cluster*

• Cluster changes are invisible to clients

• Works best with 11gR2 client (n/a on Windows)

racdb\_taf =

(DESCRIPTION =

**(ADDRESS = (PROTOCOL = TCP)(HOST = beta-scan)(PORT = 1521))**

(LOAD\_BALANCE = YES)

(CONNECT\_DATA =

(SERVER = DEDICATED)

(SERVICE\_NAME = racdb\_taf)

(FAILOVER\_MODE =

(TYPE = SELECT)(METHOD = BASIC)(RETRIES = 180)(DELAY = 5)

)

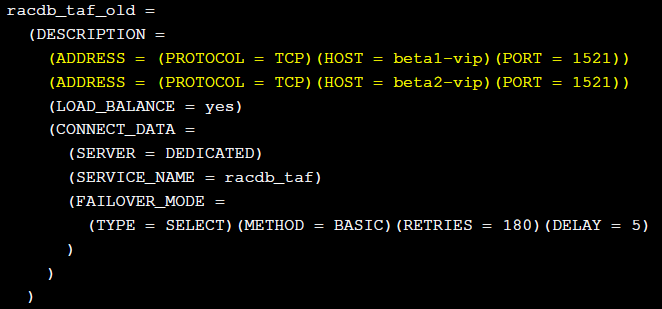
)

)

**Client Configuration**

Old way (10.2 client) still works

• No benefit of server side load balancing



**Listener Configuration**

Two Listeners

– Both run in GRID home

• Local (database) Listener

– One on each node

– Registers local instance (and ASM)

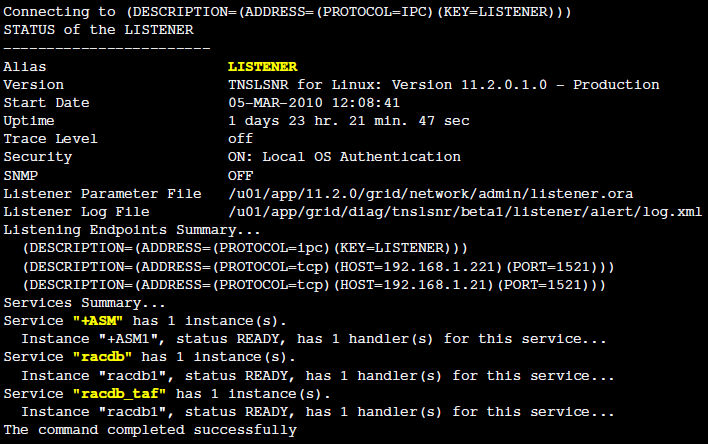
• SCAN Listener

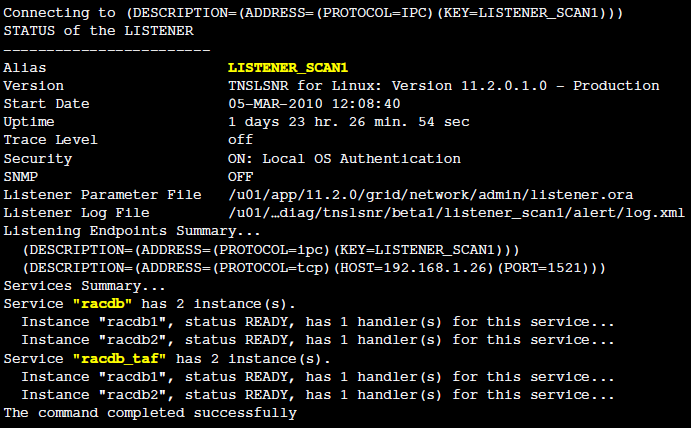
– Up to 3 per cluster

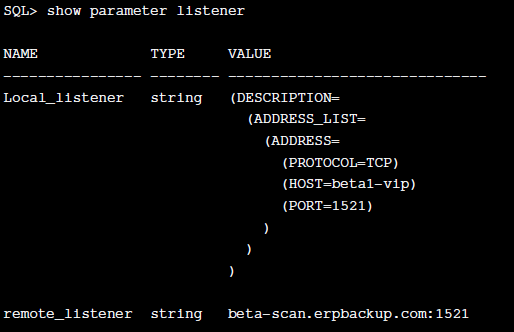
– Can migrate around cluster

– Registers all database instances & services

– Receives Load Balance Advisory







**SCAN Configuration**

Requires network configuration in place prior to install

• Requires either DNS ***or GNS (Grid Naming Service)***

• GNS requires 3 IPs acquired from DHCP

• DNS recommended for “manual” configuration

• 3 additional IPs on public network for SCAN-VIPs

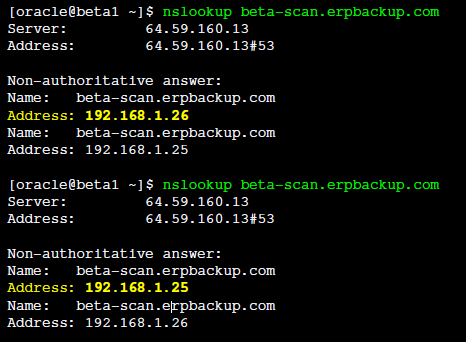
• Single SCAN hostname resolves to 3 IPs

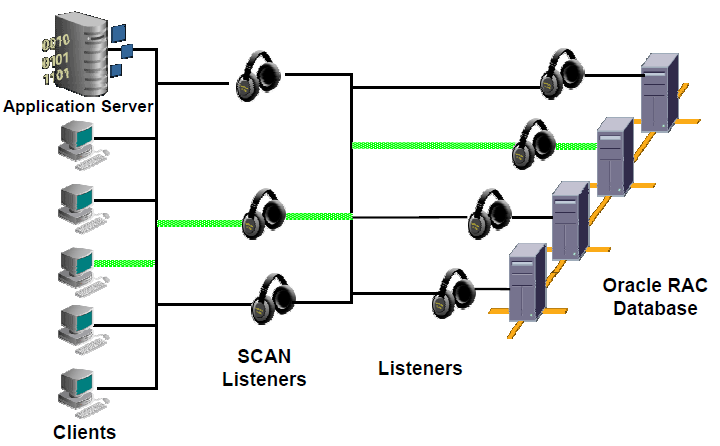
• During installation, DNS resolution provides 3 IP’s which are used to create 3 SCAN-VIP / Listener pairs scattered across the cluster

*See Metalink Doc ID Doc ID 887522.1 -*

*11gR2 Grid Infrastructure SCAN Explained*

**DNS Round Robin Configuration**





**SCAN – How it works!!**

1. Client requests DNS resolution of SCAN hostname

2. DNS responds with circulating list of 3 IPs

3. Client chooses first, or random, IP in list

4. Client connects to SCAN Listener at chosen IP

and requests database service

5. SCAN Listener is receiving service registrations and load

balance advisories from all instances in cluster

6. SCAN Listener chooses least loaded instance offering

requested service

7. SCAN Listener re-directs connection request to Local

Listener on that node

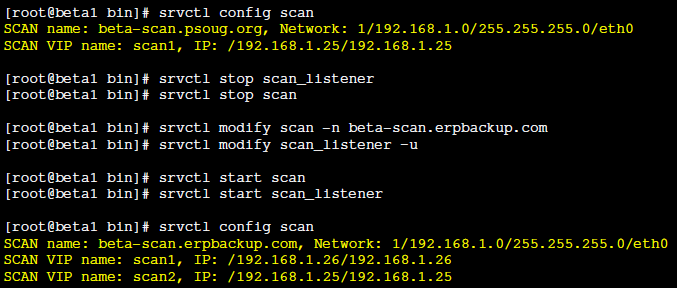
8. Local Listener accepts request and establishes session.

**SCAN Maintenance**

To re-configure later (not documented), see Metalink Doc ID

972500.1 *How to Modify SCAN Setting after Installation*

Example: Modify DNS to add additional SCAN VIP and change SCAN hostname



**Load Balancing**

**Services**

**Failover**

Failover is the process when the connection is routed to another node as the existing one is in an unavailable state.

Connect-time Failover

Supports for the failover of connections establishment request to the other surviving VIP if the selected VIP is down

fail\_over=true

After-connect Failover

Supports for the failover of user sessions to another node if the connected node goes down. This is managed by RAC. “Truly RAC”.

failed\_mode=(method =basic/preconnect)

(type=session/select)

(delay=<time latency between each retry>)

(retries=<max no. of times to retry>)

**Adding a New Node**

* Install same OS on the 3rd node.
* Configure VIP (eth0:0) on top of Public IP using ‘neat’
* Update /etc/hosts file with all node related Private IP mapping to Private Hostnames
* Configure DNS Client
* Configure Shared Storage
* Create groups, users, directory structures & assign permissions
* Install required OS/Clusterware packages
* Update kernel parameters & sync date/time on all nodes
* Update .bash\_profile with the required env variables
* Configure SSH for password-less connectivity
* Perform pre-cluster verification check
  + Component level
  + Cluster level

* Add existing Grid Infrastructure to new node
* Add existing RDBMS Home to new node
* Make changes to run the existing database on the new node.
* Update OCR using SRVCTL.

**Deleting a Node**

* Stop & unregister cluster resources belonging to the node being deleted. (This can be done from any node)
* Drop Redo Log thread & Undo tablespace
* Stop Clusterware daemons & bring down the complete clusterware
* Update Grid Home Inventory on deleted node & undeleted nodes
* Similarly, update RDBMS Home directory on deleted node & undeleted nodes
* Uninstall/Remove GI & RDBMS software from the node to be deleted using ‘deinstall’ utility
* Backup OCR

Split?

22. What is split brain syndrome?  
Ans:  
Will arise when two or more instances attempt to control a cluster database. In a two-node environment, one instance attempts to manage updates simultaneously while the other instance attempts to manage updates.

27. What is FAN?  
Ans:  
Applications can use Fast Application Notification (FAN) to enable rapid failure detection, balancing of connection pools after failures, and re-balancing of connection pools when failed components are repaired. The FAN process uses system events that Oracle publishes when cluster servers become unreachable or if network interfaces fail.  
  
28. What is FCF?  
Ans:  
Fast Connection Failover provides high availability to FAN integrated clients, such as clients that use JDBC, OCI, or ODP.NET. If you configure the client to use fast connection failover, then the client automatically subscribes to FAN events and can react to database UP and DOWN events. In response, Oracle gives the client a connection to an active instance that provides the requested database service.

29. What is TAF? and TAF policies?  
Ans:  
Transparent Application Failover (TAF) - A runtime failover for high availability environments, such as Real Application Clusters and Oracle Real Application Clusters Guard, TAF refers to the failover and re-establishment of application-to-service connections. It enables client applications to automatically reconnect to the database if the connection fails, and optionally resume a SELECT statement that was in progress. This reconnect happens automatically from within the Oracle Call Interface (OCI) library.

33. What are nodeapps?  
Ans:  
[VIP](http://satya-racdba.blogspot.com/2012/07/virtual-ip-vip-address-rac-oracle.html), listener, ONS, GSD

38. How to know the public IPs, private IPs, [VIPs](http://satya-racdba.blogspot.com/2012/07/virtual-ip-vip-address-rac-oracle.html) in RAC?  
Ans:  
# [olsnodes](http://satya-racdba.blogspot.com/2010/01/olsnodes-commands.html) -n -p -i  
node1-pub       1       node1-prv       node1-vip  
node2-pub       2       node2-prv       node2-vip

41. What is HAS (High Availability Service) and the commands?  
Ans:  
HAS includes ASM & database instance and listeners.  
[crsctl](http://satya-racdba.blogspot.com/2010/01/crsctl-commands.html)check has  
crsctl config has  
crsctl disable has  
crsctl enable has  
crsctl query has releaseversion  
crsctl query has softwareversion  
crsctl start has  
crsctl stop has [-f]

ASM?

Automatic Storage Management (ASM) is a new type of file system. ASM provided a foundation for highly efficient storage management with kernelized asynchronous I/O, direct I/O, redundancy, striping, and an easy way to manage storage.

ASM is recommended file system for [RAC](http://satya-racdba.blogspot.com/) and single instance ASM for storing database files.

This provides direct I/O to the file and performance is comparable with that provided by raw devices. Oracle creates a separate instance for this purpose.

**[ASM Instance](http://www.blogger.com/blogger.g?blogID=2511956539743992936)**

The ASM functionality is controlled by an ASM instance. This is a special instance, not a database instance, just the memory structures and as such is very small and lightweight.

* INSTANCE\_TYPE - Set to ASM. The default is RDBMS.
* ASM\_DISKGROUPS - The list of diskgroups that should be mounted by an ASM instance during instance startup, or by the ALTER DISKGROUP ALL MOUNT statement. ASM configuration changes are automatically reflected in this parameter.
* ASM\_DISKSTRING - Specifies a value that can be used to limit the disks considered for discovery. The default value is NULL allowing all suitable disks to be considered. Altering the default value may improve the speed of diskgroup mount time and the speed of adding a disk to a diskgroup. Changing the parameter to a value which prevents the discovery of already mounted disks results in an error.
* ASM\_POWER\_LIMIT -The maximum power for a rebalancing operation on an ASM instance. The valid values range from 1 (default) to 11. The higher the limit the more resources are allocated resulting in faster rebalancing operations. This value is also used as the default when the POWER clause is omitted from a rebalance operation. A value of 0 disables rebalancing.
* ASM\_PREFERRED\_READ\_FAILURE\_GROUPS - This initialization parameter value (default is NULL) is a comma-delimited list of strings that specifies the failure groups that should be preferentially read by the given instance. This parameter is generally used only for clustered ASM instances and its value can be different on different nodes. This is from [Oracle 11g](http://satya-dba.blogspot.com/2009/01/whats-new-in-11g.html).
* DB\_UNIQUE\_NAME - Specifies a globally unique name for the database. This defaults to +ASM but must be altered if you intend to run multiple ASM instances.

Normal?

* Normal redundancy - for 2-way mirroring, requiring two failure groups, when ASM allocates an extent for a normal redundancy file, ASM allocates a primary copy and a secondary copy. ASM chooses the disk on which to store the secondary copy in a different failure group other than the primary copy.
* High redundancy - for 3-way mirroring, requiring three failure groups, in this case the extent is mirrored across 3 disks.
* External redundancy - to not use ASM mirroring. This is used if you are using hardware mirroring or third party redundancy mechanism like RAID, Storage arrays.

Dg?

**Creating diskgroups**

SQL> CREATE DISKGROUP dg\_asm\_data NORMAL REDUNDANCY

FAILGROUP failure\_group\_1 DISK

'/devices/diska1' NAME diska1, '/devices/diska2' NAME diska2,

FAILGROUP failure\_group\_2 DISK

'/devices/diskb1' NAME diskb1, '/devices/diskb2' NAME diskb2;

For two-way mirroring we would expect a diskgroup to contain two failure groups, so individual files are written to two locations.

SQL> CREATE DISKGROUP dg\_asm\_fra HIGH REDUNDANCY

FAILGROUP failure\_group\_1 DISK

'/devices/diska1' NAME diska1, '/devices/diska2' NAME diska2,

FAILGROUP failure\_group\_2 DISK

'/devices/diskb1' NAME diskb1, '/devices/diskb2' NAME diskb2,

FAILGROUP failure\_group\_3 DISK

'/devices/diskc1' NAME diskc1, '/devices/diskc2' NAME diskc2;

For three-way mirroring we would expect a diskgroup to contain three failure groups, so individual files are written to three locations.

SQL> CREATE DISKGROUP dg\_grp1 EXTERNAL REDUNDANCY

DISK '/dev/d1','/dev/d2','/dev/d3','/dev/d4' ... ...;

**Listing diskgroups**

To find out all the diskgroups:

SQL> SELECT \* FROM V$ASM\_DISKGROUP;

[ASMCMD](http://satya-dba.blogspot.com/2010/02/asmcmd-10g-11g.html) equivalent for this command is [*lsdg*](http://satya-dba.blogspot.com/2010/02/asmcmd-10g-11g.html#lsdg).

**Dropping diskgroups**

Diskgroups can be deleted using the DROP DISKGROUP statement.

SQL> DROP DISKGROUP disk\_group\_1 INCLUDING CONTENTS;

SQL> DROP DISKGROUP disk\_group\_1 FORCE; ([11g R1](http://satya-dba.blogspot.com/2009/01/whats-new-in-11g.html) command)

SQL> DROP DISKGROUP disk\_group\_1 FORCE INCLUDING CONTENTS; ([11gR1](http://satya-dba.blogspot.com/2009/01/whats-new-in-11g.html) command)

**Adding disks**

We may have to add additional disks into the diskgroup to accommodate growing demand.

SQL> ALTER DISKGROUP dskgrp1 ADD DISK '/dev/d5';

SQL> ALTER DISKGROUP dg1 ADD DISK '/devices/disk\*3', '/devices/disk\*4';

**Listing client databases**

The following command shows all the database instances connected to the ASM instance.

SQL> SELECT \* FROM V$ASM\_CLIENT;

When you specify a preferred instance for a service, the service runs on that instance during standard operation. Oracle Clusterware attempts to ensure that the service always runs on all the preferred instances that have been configured for a service. If the instance fails, the service is randomly relocated to one of the available instances. You can also manually relocate the service to an available instance. If you do not specify preferred or available instances when you create a service, then by default every instance in the Oracle RAC database is a preferred instance for that service.