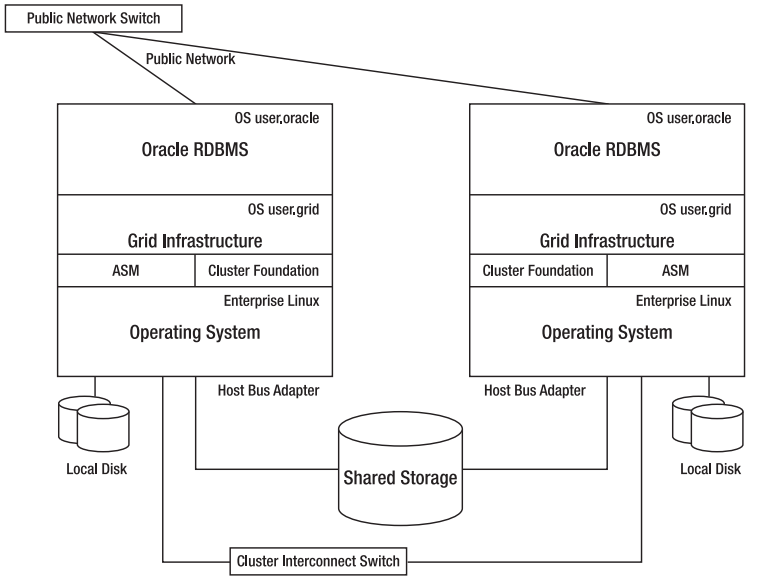
Oracle RAC

# Tổng quan về Oracle Rac

## Kiến trúc tổng thể



# Oracle Clusterware

## Oracle Clusterware concept

### What Is a Cluster?

* A group of independent, but interconnected, computers that act as a single system
* Usually deployed to increase availability and performance or to balance a dynamically changing workload

### What Is Clusterware?

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

* Allow the cluster to be managed as a whole ( single entity)
* Protect the integrity of the cluster so that data is protected and the cluster continues to function even if communication with a cluster node is severed
* Maintain a registry of resources so that their location is known across the cluster and so that dependencies between resources is maintained
* Deal with changes to the cluster such as node additions, removals, or failures
* Provide a view of resources such as network addresses and files in a file system

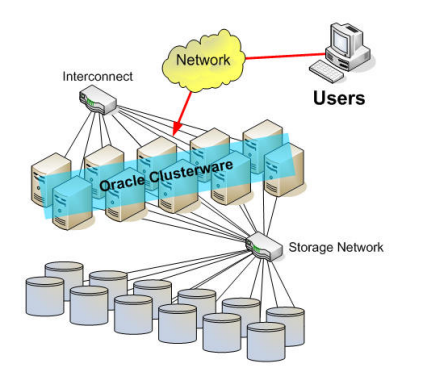
### Oracle Clusterware

* 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 Networking

* Each node must have at least two network adapters.
* Each public network adapter must support TCP/IP.
* The interconnect adapter must support:
  + User Datagram Protocol (UDP) or Reliable Data Socket (RDS) for UNIX and Linux for database communication
  + TCP for Windows platforms for database communication
* All platforms use Grid Interprocess Communication (GIPc)

### Oracle Clusterware Architecture and Services



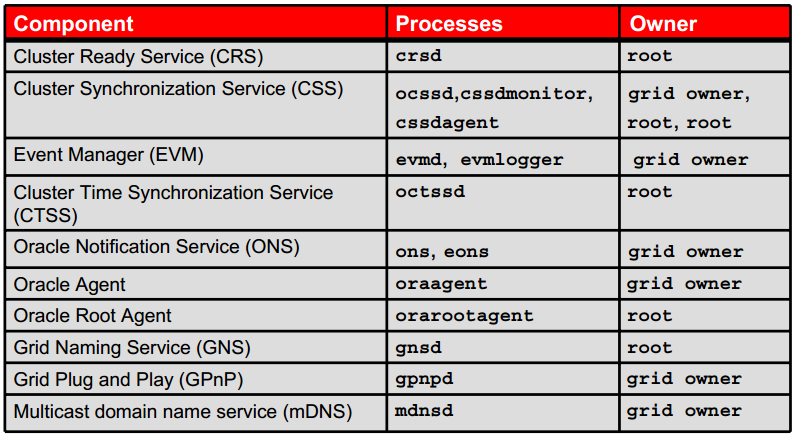
Shared disk cluster architecture supporting application load balancing and failover

Services include:

* Cluster Management, which allows cluster services and application resources to be monitored and managed from any node in the cluster
* Node Monitoring, which provides real-time information regarding which nodes are currently available and the resources they support. Cluster integrity is also protected by evicting or fencing unresponsive nodes.
* Event Services, which publishes cluster events so that applications are aware of changes in the clusterEvent services
* Time Synchronization, which synchronizes the time on all nodes of the cluster
* Network Management, which provisions and manages Virtual IP (VIP) addresses that are associated with cluster nodes or application resources to provide a consistent network identity regardless of which nodes are available. In addition, Grid Naming Service (GNS) manages network naming within the cluster.
* High Availability, which services, monitors, and restarts all other resources as required

## Oracle Clusterware Architecture

### Oracle Clusterware Process Architecture



* **Cluster Ready Service (CRS)**: Is the primary program for managing high availability operations in a cluster. The CRS process manages two types of CRS resources:
  + **Cluster resources**: A cluster resource is an Oracle Clusterware resource. Cluster resources are viewed, added, modified, or removed using the crsctl command.
  + **Local resources**: A local resource runs on every node in the cluster (no failover) and can be, for example, a listener, ASM, a disk group, or Oracle Notification Service (ONS).
* **The CRS daemon (crsd)** manages cluster resources based on configuration information that is stored in Oracle Cluster Registry (OCR) for each resource. This includes start, stop, monitor, and failover operations. The crsd process generates events when the status of a resource changes.
  + When you have Oracle RAC installed, the crsd process monitors the Oracle database instance, listener, and so on, and automatically restarts these components when a failure occurs. When a CRS resource fails, the CRS daemon attempts to restart it, if the resource is so configured. CRS fails the resource over to another node (again, if it is configured to do so) after exhausting restart attempts.
* **Cluster Synchronization Service (CSS)**: Manages the cluster configuration by controlling which nodes are members of the cluster and by notifying members when a node joins or leaves the cluster. If you are using certified third-party clusterware, then CSS processes interfaces with your clusterware to manage node membership information. CSS has three separate processes: the CSS daemon (ocssd), the CSS Agent (cssdagent), and the CSS Monitor (cssdmonitor). The cssdagent process monitors the cluster and provides input/output fencing. This service formerly was provided by Oracle Process Monitor daemon (oprocd), also known as OraFenceService on Windows. A cssdagent failure results in Oracle Clusterware restarting the node.
* **Disk Monitor daemon (diskmon)**: Monitors and performs input/output fencing for Oracle Exadata Storage Server. As Exadata storage can be added to any Oracle RAC node at any point in time, the diskmon daemon is always started when ocssd is started.
* **Event Manager (EVM)**: Is a background process that publishes Oracle Clusterware events
* **Multicast domain name service (mDNS)**: Allows DNS requests. The mDNS process is a background process on Linux and UNIX, and a service on Windows.
* **Oracle Grid Naming Service (GNS)**: Is a gateway between the cluster mDNS and external DNS servers. The GNS process performs name resolution within the cluster.
* **Oracle Notification Service (ONS):** Is a publish-and-subscribe service for communicating Fast Application Notification (FAN) events
* **oraagent**: Extends clusterware to support Oracle-specific requirements and complex resources. It runs server callout scripts when FAN events occur. This process was known as RACG in Oracle Clusterware 11g Release 1 (11.1).
* **Oracle root agent (orarootagent)**: Is a specialized oraagent process that helps CRSD manage resources owned by root, such as the network, and the Grid virtual IP address
* **Cluster kill daemon (oclskd)**: Handles instance/node evictions requests that have been escalated to CSS
* **Grid IPC daemon (gipcd)**: Is a helper daemon for the communications infrastructure
* **Grid Plug and Play(GPnP)**: reduces the costs of installing, configuring, and managing server nodes by using a Grid Naming Service within the cluster to allow each node to perform the following tasks dynamically

### Single Client Access Name

* The single client access name (SCAN) is the address used by clients connecting to the cluster.
* The SCAN provides a stable, highly available name for clients to use, independent of the nodes that make up the cluster.

### Grid Infrastructure Concept

* Cluster configuration information is maintained in the OCR.
* Only one of the **CRSD** processes actually reads from and writes to the OCR file on shared storage This process is responsible for refreshing its own local cache, as well as the OCR cache on other nodes in the cluster. For queries against the cluster repository, the OCR clients communicate directly with the local CRS daemon (CRSD) process on the node from which they originate. When clients need to update the OCR, they communicate through their local CRSD process to the CRSD process that is performing input/output (I/O) for writing to the repository on disk.
* CSS is the service that determines which nodes in the cluster are available and provides cluster group membership and simple locking services to other processes. CSS typically determines node availability via communication through a dedicated private network with a voting disk used as a secondary communication mechanism

## ASM Architecture

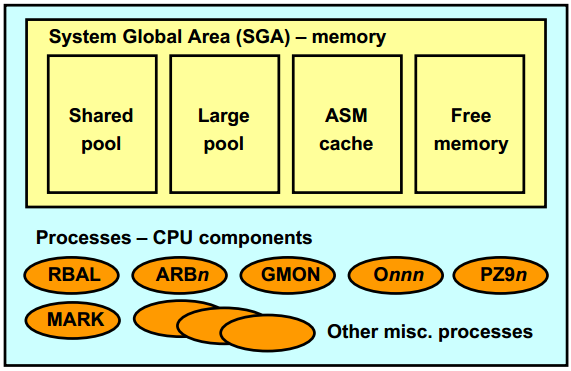
### What Is Oracle ASM?

* In essence, Oracle Automatic Storage Management (ASM) is a volume manager and file system built into the Oracle Database server.
* Raw disk volumes are allocated to ASM for management and control in the same way that raw volumes are managed by a volume manager.
* ASM is highly integrated with, and highly optimized for, the Oracle Database. It has become the best practice standard for Oracle Database storage.

### ASM Key Features and Benefits

* Stripes files rather than logical volumes
* Provides redundancy on a file basis
* Enables online disk reconfiguration and dynamic rebalancing
* Reduces the time significantly to resynchronize a transient failure by tracking changes while disk is offline
* Provides adjustable rebalancing speed
* Is cluster-aware
* Supports reading from mirrored copy instead of primary copy for extended clusters
* Is automatically installed as part of the Grid Infrastructure

### ASM Instance

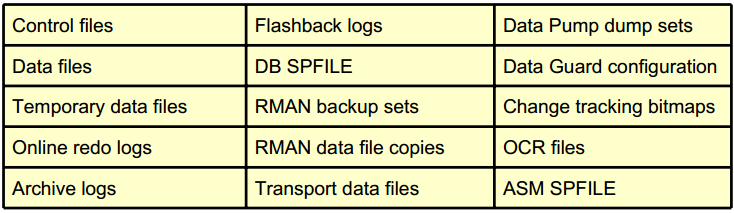


Component:

* **Shared Pool:** Used for metadata information
* **Large Pool:** Used for parallel operations
* **ASM Cache:** Used for reading and writing blocks during rebalance operations
* **Free Memory:** Unallocated memory available
* **Backgroud process:**
  + ARCn: The archiver processes
  + CKPT: The checkpoint process
  + DBWn: The database writer processes
  + DIAG: The diagnosability process
  + Jnnn: Job queue processes
  + LGWR: The log writer process
  + PMON: The process monitor process
  + PSP0: The process spawner process
  + QMNn: The queue monitor processes
  + RECO: The recoverer process
  + SMON: The system monitor process
  + VKTM: The virtual keeper of time process
  + MMAN: The memory manager process

### ASM File

ASM files are a limited set of file types stored in an ASM disk group:



* Are stored as a set or collection of data extents
* Are striped across all disks in a disk group
* Use names that begin with a plus sign (+), which are automatically generated or from user-defined aliases

+group/dbname/file\_type/file\_type\_tag.file.incarnation

### ASM Intelligent Data Placement

Places files in hot or cold regions based on:

* File attribute
* Template

Improves performance by:

* Placing frequently accessed files in the hot region

Improves disk usage by:

* Placing infrequently used files in the cold region

IDP attribute changes to existing data will be implemented by the next rebalance operation.

Intelligent Data Placement works best for:

* Data files that are accessed at different rates
* Disk groups more than 50% full
* Disks with the geometry exposed to ASM (JBOD)

Intelligent Data Placement requires disk group attributes.

* COMPATIBLE.ASM = 11.2.0 or higher
* COMPATIBLE.RDBMS = 11.2.0 or higher

Assigning Files to Disk Regions

* Create a file template:

ALTER DISKGROUP data ADD TEMPLATE datafile\_hot

ATTRIBUTES ( HOT MIRRORHOT);

* Alter the file attributes:

ALTER DISKGROUP data MODIFY FILE

'+data/orcl/datafile/users.259.679156903'

ATTRIBUTES (COLD MIRRORCOLD);

# Administering Oracle Clusterware

\*\*\*Sumarry

CRS= OCR+ Voting disk

* Voting disk: crsctl: voting disk is a file that manages information about node membership
* OCR: srvctl : OCR is a file that manages cluster and Oracle RAC database configuration information

## Managing Oracle Clusterware

### Resource Management Options

Oracle Database 11g Release 2 provides two styles of resource management for RAC databases:

* Administrator-managed configuration:
  + Uses a subpool of the built-in GENERIC server pool
  + Places resources manually as defined by the database administrator
  + Is the management strategy used in previous releases
  + Is good for a small number of nodes
* Policy-managed configuration:
  + Divides servers in a cluster into sets of server pools
  + Controls placement of all services by using server pools
  + Is Workload Management compliant
  + Is best for a large number of nodes

### Server Pools

Server pools are logical divisions of a cluster into pools of servers or nodes. They:

* Distribute a uniform workload over several servers in the cluster
* Are allocated to host databases or other resources
* Are managed using the crsctl and srvctl commands
* Support parent-child relationships among server pools
  + Top-level pools are mutually exclusive.
* Include two built-in server pools at Oracle Clusterware installation:
  + FREE: For servers that are not assigned to other pools
  + GENERIC: For administrator-managed fixed configuration and pre-11g Release 2 databases

### Crsctl and srvctl commands

* crsctl manages clusterware-related operations:
  + Starting and stopping Oracle Clusterware
  + Enabling and disabling Oracle Clusterware daemons
  + Registering cluster resources
* srvctl manages Oracle resource–related operations:
  + Starting and stopping database instances and services

## Controlling Oracle Clusterware

* When server start, crs is is automatically started by the /etc/init.d/ohasd startup script
* **Check status:**

# 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

# crsctl check cluster

* **Start /stop – enable/disable ở 1 node**

# crsctl start crs  
# crsctl stop crs

# crsctl enable crs  
# crsctl disable crs

* **To determine the location of the voting disk:**

# crsctl query css votedisk

## STATE File Universal Id File Name Disk group  
-- ----- ----------------- ---------- ----------  
1. ONLINE 8c2e45d734c64f8abf9f136990f3daf8 (ASMDISK01) [DATA]  
2. ONLINE 99bc153df3b84fb4bf071d916089fd4a (ASMDISK02) [DATA]  
3. ONLINE 0b090b6b19154fc1bf5913bc70340921 (ASMDISK03) [DATA]  
Located 3 voting disk(s).

* **To determine the location of the OCR:**

$ cat /etc/oracle/ocr.loc

ocrconfig\_loc=+DATA

local\_only=FALSE

$ ocrcheck

Status of Oracle Cluster Registry is as follows :

Version : 3

Total space (kbytes) : 262120

Used space (kbytes) : 3856

Available space (kbytes) : 258264

ID : 12663931

Device/File Name : +OCR\_VOTE

Device/File integrity check succeeded

Cluster registry integrity check succeeded

Logical corruption check bypassed due to non-privileged user

* **Check the ocssd.log for voting disks issues.**

$ grep voting <grid\_home>/log/<hostname>/cssd/ocssd.log

* **To add or delete one or more voting disks to non-ASM storage:**

# crsctl add css votedisk path\_to\_new\_voting\_disk

# crsctl delete css votedisk path\_to\_old\_voting\_disk

* **To add a voting disk to ASM:**

# crsctl replace votedisk +asm\_disk\_group

* **To migrate voting disks from non-ASM storage devices to ASM or vice versa, specify the ASM disk group name or path to the non-ASM storage device:**

# crsctl replace votedisk {+asm\_disk\_group | path\_to\_voting\_disk}

* **Locating the OCR Automatic Backups**
  + The OCR is backed up automatically.
  + Only one node performs the backup.
  + To determine the node and location of the backup:

$ ocrconfig -showbackup auto

ttdl-coredb1a 2015/11/24 15:12:50 /u01/app/11.2.0/grid/cdata/ttdl-coredb1/backup00.ocr

ttdl-coredb1a 2015/11/24 11:12:49 /u01/app/11.2.0/grid/cdata/ttdl-coredb1/backup01.ocr

ttdl-coredb1a 2015/11/24 07:12:48 /u01/app/11.2.0/grid/cdata/ttdl-coredb1/backup02.ocr

ttdl-coredb1a 2015/11/23 15:12:46 /u01/app/11.2.0/grid/cdata/ttdl-coredb1/day.ocr

ttdl-coredb1b 2015/11/13 14:56:59 /u01/app/11.2.0/grid/cdata/ttdl-coredb1/week.ocr

* + Files could be spread across nodes due to outages.
  + The backup frequency and retention policies are:
    - Every four hours: CRS keeps the last three copies.
    - At the end of every day: CRS keeps the last two copies.
    - At the end of every week: CRS keeps the last two copies.
  + Default directory:

[grid@...#](mailto:grid@...) $ORACLE\_HOME/cdata/<cluster name>

* **Adding, Replacing, and Repairing OCR Locations**
  + Add an OCR location to either ASM or other storage device:

# ocrconfig -add +DATA2

# ocrconfig -add /dev/sde1

* + To replace the current OCR location:

# ocrconfig -replace /dev/sde1 -replacement +DATA2

* + To repair OCR configuration, run this command on the node on which you have stopped Oracle Clusterware:

# ocrconfig -replace /dev/sde1 -replacement +DATA2

* **Removing an Oracle Cluster Registry Location**

# ocrconfig -delete +DATA2  
# ocrconfig -delete /dev/sde1

* **Migrating OCR Locations to ASM**
  + Ensure that Oracle Clusterware is upgraded to 11g Release 2

$ crsctl query crs activeversion  
Oracle Clusterware active version on cluster is [11.2.0.1.0]

* + Start ASM on all nodes and create a disk group that has at least 1 GB of space and has at least normal redundancy.
  + To add an OCR location to an ASM disk group, run the following command as root:

*#* ocrconfig -add +DATA2

* + To remove storage configurations no longer in use, run the following command as root:

# ocrconfig -delete /dev/raw/raw1

# ocrconfig -delete /dev/raw/raw2

* **To perform a backup:**
  + Physical backup

# ocrconfig -manualbackup

* + To display a list of manual backups:

$ ocrconfig –showbackup manual

host02 2009/07/28 16:59:17

/u01/app/.../cdata/cluster01/backup\_20090728\_165917.ocr

* + To perform a logical backup:

# ocrconfig -export /home/oracle/ocr.backup

* **Recovering the OCR by Using Physical Backups**
  + Locate a physical backup:

$ ocrconfig –showbackup

* + Stop CRS on all nodes:

# crsctl stop crs -f

* + Restore the physical or logical OCR backup:

# ocrconfig –restore /u01/app/.../cdata/cluster01/day.ocr

or

# ocrconfig –import /shared/export/ocrback.dmp

* + Restart Oracle High Availability Services on all nodes:

# crsctl start crs

* + Check status

# ocrcheck

## Administrator-managed vs Policy managed

* Administrator-managed: Database administrators define on which servers a database resource should run, and place resources manually as needed. This is the management strategy used in previous releases
* Policy managed: Database administrators specify in which server pool (excluding generic or free) the database resource will run. Oracle Clusterware is responsible for placing the database resource on a server.

# RAC Database Monitoring and Tuning

## Most Common RAC Tuning Tips

Application tuning is often the most beneficial!

* Reduce long full-table scans in OLTP systems.
* Use Automatic Segment Space Management (ASSM).
* Increase sequence caches.
* Use partitioning to reduce interinstance traffic.
* Avoid unnecessary parsing.
* Minimize locking usage.
* Remove unselective indexes.
* Configure interconnect properly.

## Config for high avaibility

