(6)

Clusterware Architecture :-

→ The 2 important files of oracle clusterware are

(1) OCR file

(2) voting file

The above 2 files needs to be placed in the shared storage either in raw partitions or cluster file system.

→ In 11gR2, we can place these files in ASM diskgroups.

OCR File (Oracle Cluster Registry) :-

OCR file contains entire cluster configuration information like node names, IP's info, databases, instances, services, listeners etc. that are registered with the cluster.

→ CRSD (cluster Ready Service Deamon) maintains information in OCR file

→ Various clients of OCR are

(1) SRVCTL

(2) DBCA

(3) vipca

(4) Netca

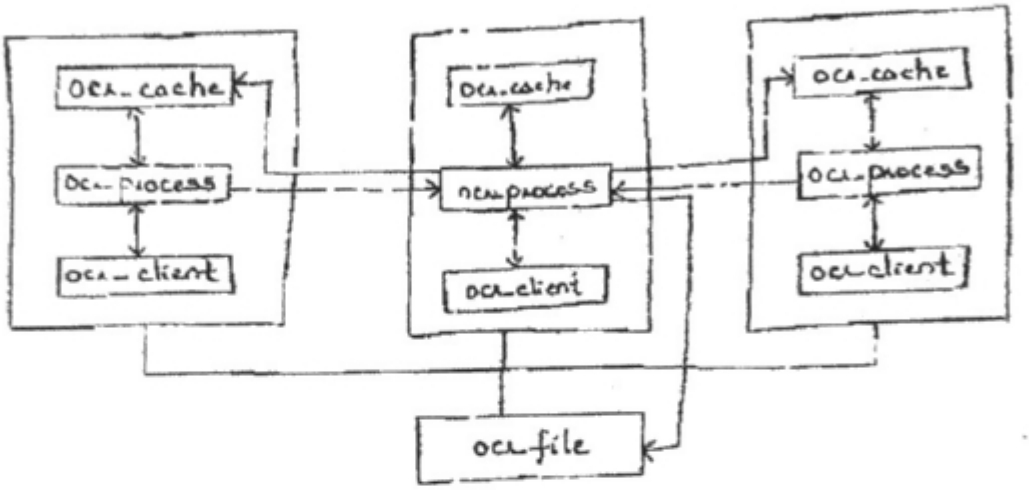
(5) Enterprise manager console

→ Inorder to optimize the performance of queries on OCR, oracle RAC follows distributed OCR cache architecture.

→ Every node that is participating in cluster maintains IN MEMORY copy of OCR file which is called OCR cache.

→ out of all the nodes, one node will be designated as the master node and master node takes the responsibility of updating local OC cache and remote OCR caches

→ At any point of time, only master node OCR process will be in directly contact with the OCR file



→ By default, master node takes the backup of OCR file for every 4 hours, at the end of the day and at the end of the week.

→ To know the master node and the default backup location execute the following command

$ocrconfig -showbackup

→ If the master node goes down or if the master node gets evicted from the cluster, any one of the surviving node becomes the master

→ In 11g Rel2, every node that is participating in cluster maintains OLR in the local hard disk.

→ ohasd (oracle high availability service daemon) which was introduced in 11g Rel2 maintains information in OLR (Oracle local Registry)

Voting File/voting Disk :-

→ This is something like quorum disk in 9i RAC or any 3rd party clustering technology

→ Every node participating in cluster has to vote in voting disk for the specified default time interval.

→ The main purpose of voting disk is to check the heartbeat mechanism.

→ if any one of the node fails to vote in the voting disk the other node considers it as dead node.

* → Inorder to avoid split brain syndrome (every clustering technology has), oracle RAC majority voting principle, because of this reason Oracle RAC maintains odd no of voting lists.

	No: of disks/voting disks
External Redundancy	1
Normal Redundancy	3
High Redundancy	5

→ In a RAC system, every node that is participating in cluster has 2 types of heart beats

(1) N/w heartbeat

(2) Disk heartbeat

→ If any of the heartbeat fails, node gets evicted from the cluster.

Oracle RAC Architecture and Internals :-

Has
Cache Fusion

In version 9i, oracle has made lots of architectural changes to its high availability clustering technology & renamed the technology from ops (oracle Parallel Server) to Real Applicat clusters

"Cache fusion is nothing but movement of data or transfer of data across the instances through cluster interconnect or private interconnect is called cache fusion"

Cache fusion was introduced in version 9i of oracle

Cache coherence :-

Maintaining consistency of data across all the instances i a RAC system is called cache coherence

Resource Affinity :-

Dynamically changing the ownership of resources in RAC system is called resource affinity

GRD (Global Resource Directory)

Every node that is participating in cluster maintains GRD.

some portion of memory from shared pool will be

allocated in GRD GRD is maintained by

① GCS (Global cache Service)

② GES (Global Enqueue Service)

GRD contains

① Data Block Address

② Role of the resource

③ Mode of the resource

④ Location of latest version of the resource

→ Role of the resource could be local or global

→ In a standalone system, role of the resource is always local to that particular instance

→ If RAC is enabled role of the resource could be either local or global

→ Mode of the resource could be NULL (or) shared (or) exclusive

NULL :-

→ NULL represents a place holder on the resource without much significance

SHARED :-

→ Generally resources are held in shared mode incase of select

EXCLUSIVE :-

→ Generally resources are held in exclusive mode incase of update.

→ In a RAC system, this daemon is going to start, stop, and monitor the states of the resources that are registered with the cluster.

→ This daemon gets restarted automatically if it fails

→ This is the only daemon which runs under root account

CSSD (Clustery Synchronization Service Daemon) :-

→ This daemon also comes into the picture in standalone system if the database storage area location is ASM.

→ In 10g, inorder to start this daemon we need to execute the following script as root user before starting ASM instance

```
#cd $ORACLE_HOME/bin
#./localconfig add
```

→ In a RAC system this daemon provides node membership for every node that is participating in clustery.

→ This daemon notifies all other members in the clustery whenever a node joins or leaves the clustery.

→ Failure to start this daemon will not allow a node to join the clustery.

CSSD Monitor :-

→ Monitors the scheduling of cpu. If it detects the hung state of a node, it is going to reset the node to maintain data integrity. This process is called Stonith fencing. Prior to 10.2.0.4 this task was carried out by OprocD (Oracle Process Monitor Daemon)

EVMB (Event Monitor Daemon) :-

→ Scans call out directory and invokes callouts w.r.t the events that are detected.

To disable the cluster

```
$ crsctl disable crs
```

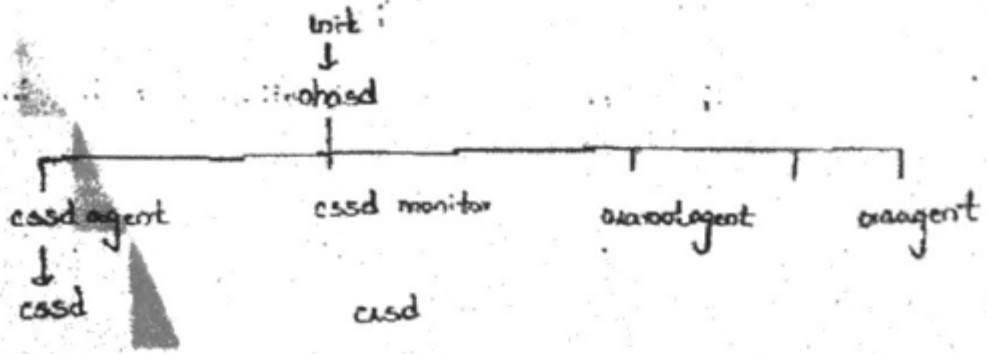
To enable the cluster

```
$ crsctl enable crs
```

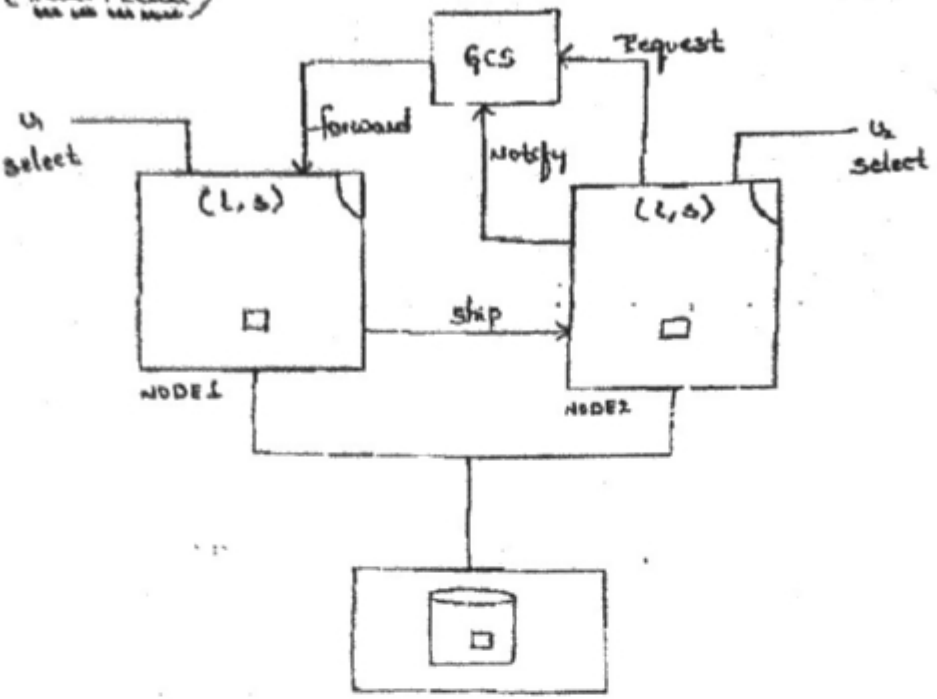
Cluster startup process in HpTru64 :-

inittab ---> inittab.orig

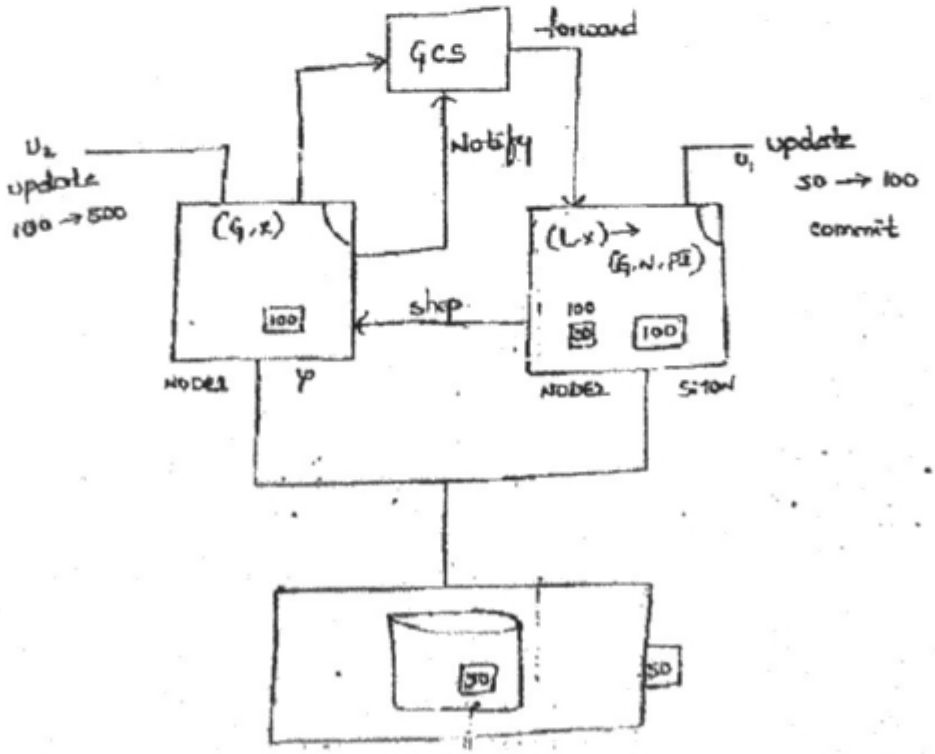
3 5 /etc/init.d/init.ohasd



Scenario-1
(Read / Read)



Scenario-II
(Write / write)

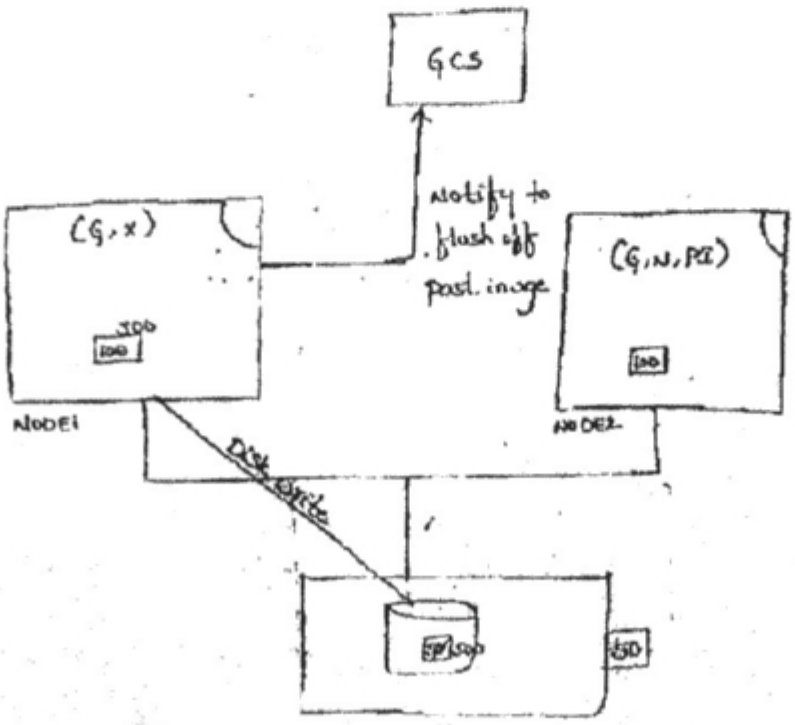


Concept of past image :-

→ Past image concept was introduced in 9i version of oracle. When an instance requests a resource which is already available in some other instance (updated version), the owning instance retains a copy of the resource before shipping onto the requested instance. This copy of the resource is called past image. Past images are generally held in null mode and they are used to reconstruct the version of the resource in the event of requested instance failures.

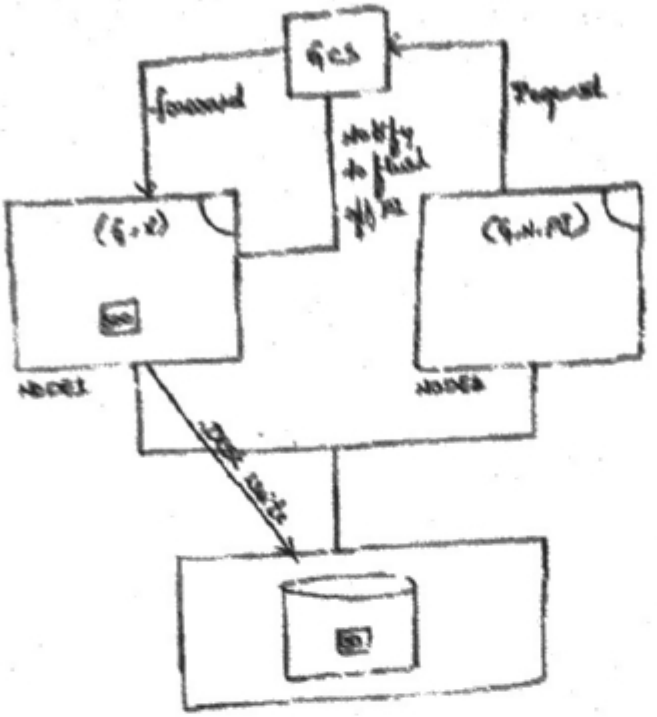
SCENARIO - III

Disk Write : Instance 1 wants to write



Scenario - III

Disk write : Instance 2 wants to write.



NOTE :- In a 2 node cluster an instance gets a resource in 2 hops

In 'n' no. of node cluster (except 2 node) an instance gets a resource in max 3 hops