**1. Why we are using vip in RAC? Because before 10g, there were no concept of vip.**

If user connected to the INSTANCE  using physical ip, and if the node goes down, then there is no way for the user to know whether node is available or not. So it need to wait for a long time, till it gets timed out by the network.

However If i use logical vip( on top of physical vip), then if node goes down, then CRS will failover this vip to other surviving node. And the user will get the connection error quickly( like TNS no listener available ).

**61. How can i get the cluster name in RAC?**

olsnodes -c

**38. What is GES and GCS?**

GES and GCS are two important parts of GRD(Global resource Directiory)

GES and GCS have a memory structure in Global resource , which is distributed across the instance. It is part stored in shared pool section.

Global enqueue service ( GES) performs concurrency control  on dictonary cache, library cache locks and transactional locks. This mechanism ensures that all the instances in cluster, know the locking status of each other . i.e If node 1 want to lock a table , then it need to know what type of lock is present in other node.  Background processes like **LCK0, LMD and LMON** .

Glocal Cache Service(GCS) handles the block management. It maintains  and tracks the location and status of blocks.  **It is responsible for block transfer across instances.**  LMS is primary background process .

**62. What is disktimeout and miscount ?**

**Miscount** is the maximum delay in network heartbeat. By default miscount is set to 30 second . If the nodes are unable to communicate with each other through private interconnect for 30 seconds( miscount value), then node eviction will be initiated.

**Disktimeout**: It is the maximum voting disk heartbeat delay. Default is 200 seconds. If the CSSD is unable to write more than half of voting disks, then eviction will happen.

**2. If i have a 8 node RAC, then how many scan listeners are required?**

3 SCAN listeners sufficient for any RAC setup. It is not mandatory for scan listener to run on all the nodes.

**Total IP’s : (n\*3)+3**

**3. How SCAN knows which node has least load?**

Load balance Advisory provides load information to scan.

**8. What are some RAC specific parameters ?**

* **undo\_tablespaces**
* **cluster\_database**
* **cluster\_database\_instances**
* **cluster\_interconnects**
* **remote\_listener**
* **thread**

**5. What current block and CR block and PI in RAC?**

Data block requests from global cache are of two types.

**current block(cur)** – > When we want to update a data, oracle must locate the most recent version of the block in cache, it is known as current block

**consistent read(CR) –** > When we want to read a data, then only committed data will be provided( with help of undo). that is known as consistent read.

**Past image(PI)** – When  node A wants to updates the block, which is present on node B and node B has also updated the block , then node B will send the current copy of the block to Node A, it will keep a past image( PI)  of the block , until it is written to the disk. Once commit happens on node B for that transaction or when checkpoint happens, the PI images will be flushed to disk.

There can be multiple CR blocks, But there will be always one Current block.

There can multiple scur(shared current) , But only xcur( exclusive current).

**6. What is gc buffer busy wait?**

Mean a session is trying to access a buffer in buffer cache, But that particular buffer is currently busy with global cache operation.

So during that time gc buffer busy wait will happen.

Example –

* Let’s say session A want to access block id 100 , But currently that block is in buffer cache of session B.
* So session A requested session B LMS process to transfer the block.
* While transfer is going on , session B also tried to access that block. But as already that block/buffer is already busy in global cache operation. Session B has to wait with wait event, gc buffer busy wait.

Reasons – Concurrency related, Right hand index growth.

other reason might be lack of cpu, slow interconnect .

**54. How can we improve global cache performance?**

1. We can increase the number of LMS processes, by increasing gc\_server\_process.
2. We can set “\_high\_priority\_processes”=”LMS\*|LGWR\*”

**55. What is gc block lost?**It indicates issue with interconnect. If a requested block is not received by the instance in 0.5 seconds, the block is considered to be lost.

**9. Why RAC has separate redo thread for each node?**

In RAC, each instance have their own lgwr process , So there has to be separate online redolog for each instance ( called as thread), So that lgwr will write to the respective redo log.

**10. Why RAC has separate undo tablespace for each node?**

If we keep only one undo, then it need more coordination between nodes and it will impact the traffic between the instances.

**11. Explain about local\_listener and remote\_listener parameter in RAC?**

In RAC, local\_listener parameter points to node vip and remote\_listener is set to the scan

Purpose of Remote Listener is to connect all instances with all listeners so the instances can propagate their load balance advisories to all listeners. Listener uses the advisories to decide which instance should service client request. If listener get to know from advisories that its local instance is least loaded and should service client request then listener passes client request to local instance. If local instance is over loaded then listener can use TNS redirect to redirect client request to a less loaded instance means remote instance. This Phenomenon is also called as **Server Side Load balancing.**

**15. What is TAF?**

TAF provides run time failover of connection. There are different options we can mention while creating taf policy.

Let’s say we created TAF with select option. Now Suppose a user connecting to using the taf and running a select statement. While select statement is running, the node on which the select statement  running crashed. So the select statement will be transparently failed over to other node and select statement will be completed and results will be fetched.

**41. What is TAF in oracle RAC?**

1. BASIC
2. PRECONNECT
3. SELECT  FAILOVER
4. SESSION FAILOVER

**21. Explain different ways to find master node in oracle rac?**

 Grep occsd Log file. [oracle @ tadrac1]: /u1/app/../cssd >grep -i “master node” ocssd.log | tail -1. …

1. Grep crsd log file. [oracle @ tadrac1]: /u1/app/../crsd>grep MASTER crsd.log | tail -1.
2. Query V$GES\_RESOURCE view.
3. ocrconfig -showbackup. The node that store OCR backups is the master node.

**26. My clusterware version is 11gr2 , can i install oracle 12c database? is the viceversa possible( means clusteware version 12c and oracle database version 11g?)?**

My clusterware version can be same or higher than the the database version. But a 12c database will not work on 11g grid.

**33. What is GPNP profile?**

Grid plug and play(GPNP) file is small xml file present at os local file system . Each node has their owner GPNP file.

GPNP file is managed by GPNP daemon.

It stores information like asm disk string , asm spfile which are required to start the cluster.

– **Storage to be used for CSS**

– **Storage to be used for ASM : SPFILE location,ASM DiskString**

– public private network details.

When clusteware is started, It needs voting disk( which is inside ASM). So first it will check the gpnp profile to get the voting disk location( asm\_diskstring is defined inside gpnp profile) .As asm is not up at this point, asm voting disk file will read using kfed read command. ( We can run kfed, even when asm instance is down).

**34. What are the software stacks in oracle clusterware?**

From 11g onward, there are two stacks for clusterware is CRS.

1. lower stack is high availability cluster service stack ( managed by ohasd daemon)
2. upper stack is CRSD stack ( managed by CRSd daemon)

**39. What is dynamic remastering?**

Mastering of a block means, master instance monitors the state of blocks  until the remastering happens due of few of the scenarios like instance crash etc.

GRD stores useful information like data block address, block status, lock information, scn, past image etc. Each instance have some of the GRD data in their SGA. i.e. any instance which is master of the block or resource, will maintain the GRD of that resource in their SGA.

Mastering of a resource is decided based on the demand. If a particular resource is mostly accessed from node 1, then node1 will become the master of that resource. And if after some time if node 2 is heavily accessing the same resource, then all the resource information will be moved the node2 GRD.

LMON, LMD, LMS are responsible for dynamic remastering.

Remastering can happen due to below scenarios.

1. Resource affinity – > GCS keeps tracks of the number of GCS  request per instance and per objects . If one instance is heavily accessing the object blocks, compare to other nodes, Then gcs can take decision to migration all the object resource to the heavily accessed instance.
2. Manually remastering – > We can manually remaster a object
3. Instance crash – > If instance is crashed, then its GRD data will be remastering to the existing instances in cluster.

**74. Can we see DRM( Dynamic Resource Mastering) related information in oracle RAC?**

Yes we can see DRM related data in  **gv$gcspfmaster\_info***by passing the object\_id.*

**42. Can we have multiple SCAN(name) in a RAC?**

From 12c onwards, We can have multiple scan with different subnets. As part of installation only scan will be configured. Post installation we need to configure another SCAN with different subnet( If required).

**43. In RAC, where we define the SCAN?**

We can define SCAN with below 2 option.

1. Using corporate DNS
2. Using Oracle GNS( Grid naming service)

**50. What is HAIP?**

HAIP, High Availability IP, is the Oracle based solution for load balancing and failover for private interconnect traffic. Typically, Host based solutions such as Bonding (Linux)is used to implement high availability solutions for private interconnect traffic. But, HAIP is an Oracle solution for high availability.

Essentially, even if one of the physical interface is offline, private interconnect traffic can be routed through the other available physical interface. This leads to highly available architecture for private interconnect traffic.

The ora.cluster\_interconnect.haip resource will pick up a  highly available virtual IP (the HAIP) from “link-local” (Linux/Unix)  IP range (169.254.0.0 ) and assign to each private network.   With HAIP, by default, interconnect traffic will be load balanced across all active interconnect interfaces. If a private interconnect interface fails or becomes non-communicative, then Clusterware transparently moves the corresponding HAIP address to one of the remaining functional interfaces.

$ crsctl stat res ora.cluster\_interconnect.haip -init

NAME=ora.cluster\_interconnect.haip

TYPE=ora.haip.type

TARGET=ONLINE STATE=ONLINE on dbhost1

 Here if you , while installing, we have given private interrconnect as 192.168.1.0 ( ens225) , But while starting the cluster, a new vip as 169.254\* has been assigned, so gv$cluster\_interconnect shows ip\_address as 169.254\*.

NOTE – For the HAIP, to failover to other interconnect, there has to be another physical interconnect,

**52. What is rebootless node fencing?**

Prior to 11.2.0.2  , If failures happens with RAC components like  private interconnect and  voting disk accessibility, then to avoid split brain , oracle clusterware does fast reboot of the node  But the problem was that node reboot that, if any non cluster related processes are running on node, then those also gets aborted. Also , with reboot, the resources also need to be remasterd, which is expensive sometime.

Also if sometime if some issue or blockages in the i/o temporarily then also clusterware will misjudge that, initiate reboot.

So to avoid this, from 11.2.0.2 onward, this method has been improved, and known as reboot-less node fencing.

1. First clusterware finds which node to be evicted
2. Then i/0 generating processes will be killed on the problematic node.
3. Clusterware resources will be stopped on the problematic node
4. OHASD process would be running, will try continuously to start CRS, till issue is resolved.

But if due to any issue, the it is unable to stop the processes on the problematics node( i.e rebootless fencing fails) , then fast reboot will be initiated by cssd.

**58. Where the OLR is stored? When olr backup is created.**

By default, OLR is located at Grid\_home/cdata/host\_name.olr

The OLR is backed up after an installation or an upgrade. After that time, you can only manually back up the OLR. Automatic backups are not supported for the OLR.

**60. Why we need odd number of voting disks in RAC?**

A node must be able to access strictly more than half of the voting disks at any time. So if you want to be able to tolerate a failure of n voting disks, you must have at least 2n+1 configured. (n=1 means 3 voting disks).

So whether you have 3 disks or 4 disks. only failure of 1 disks will be tolerated

**67. asm spfile location is missing inside gpnp profile, Then how will asm instance startup?**

For this, we need to understand the search order of asm spfile

1. First it will check for asm spfile location inside gpnp profile
2. If no entry is found inside gpnp profile, then it will check the default path $ORACLE\_HOME/dbs/spfile+ASM.ora or a pfile.

**68. How you find out issue with private interconnect?**

You can use traceroute to check if any issue with data transfer in private interconnect.

**69. How to apply patch manually in RAC?**

First do you the patch conflict check against the OH.

Then rootcrs.sh -prepatch to unlock the crs ( without unlocking the crs, opatch utility cannot do any modification to the grid home)

Then opatch apply ( to apply the patch)

Then rootcr.sh -postpatch to lock the crs

**70.Lets say, you applied patch on node 2, and ran rootcrs.sh -post , and now it shows patch mismatch. But  when you checked the oracle inventory(opatch lsinventory), Patches are same across both the nodes. Then what you will do?**

In this case, you can run kfod command to find the missing patch.

**Action plan:**

Run kfod op=PATCHES on all the nodes and see on which nodes if any patch is missing.

Lets say you found that patch 45372828 is missing on node 2, then

On node2 as a root user , run below command

root#$GRID\_HOME/bin/patchgen commit 45372828

After that you can run  below commands to verify  whether patch level is same or not .

kfod op=PATCHLVL  
kfod op=PATCHES

After the confirmation , you can run rootcrs.sh -patch

**72.In a 12c two node RAC, What will happen, if I unplug the network cable for private interconnect?**

Rebootless node fencing will happen. i.e the node which is going to be evicted, on that node all cluster services will be down. and the services will be moved to the surviving node. And crs will do the restart attempt continuously until the private interconnect issues fixed.  Please note – the node will not be reboot, only the cluster services willl go down.

However Prior to 11.2 , In this situation, the node reboot will occur.

**73.In a rac system , What will happen if i kill the pmon process?**

The pmon will be restarted automatically.

**74. Can we see DRM( Dynamic Resource Mastering) related information in oracle RAC?**

Yes we can see DRM related data in  **gv$gcspfmaster\_info***by passing the object\_id.*

**75. What is Grid infrastructure Management Repository(GIMR)?**

Grid Infrastructure Management Repository (GIMR) is a centralised infrastructure database for diagnostic and performance data and resides in Oracle GI Home. It is a single instance CDB with a single PDB and includes partitioning (for data lifecycle management).

**77. If MGMTDB is not coming up for any reason, then what will be the impact on the existing databases?**

No impact on existing database, it will just give warning.

**78. What is the maximum number of voting disks we can configure?**

We can configure **upto 15 voting disk**.( from 11g onward)

**79. What is node weightage?**

Prior to 12cR2 , during node eviction,  node with lower number ( i.e which node joined the cluster first) survives .

But in 12cR2, node weightage concept has been introduced. I.e the node having more number of services or workload will survive the eviction.

There is another option to assign weightage to the services/databases using  -css\_critical=yes in srvctl database/service .

**80. Suppose someone has changed the permission of files inside grid\_home. How you will fix it?**

You can run rootcrs.sh -init command to revert the permission.

# cd <GRID\_HOME>/crs/install/  
# ./rootcrs.sh -init

Alternatively you can check the below files under $GRID\_HOME>/crs/utl/<hostname>/

– crsconfig\_dirs which has all directories listed in <GRID\_HOME> and their permissions

– crsconfig\_fileperms which has list of files and their permissions and locations in <GRID\_HOME>.

**86. What will happen if I kill the database archiver process in oracle rac node?**

It will be restarted.

**88. What is this recovery buddy feature in oracle 19c?**

Usually when instance is crashed in RAC, then one node is elected among the surviving nodes, for doing the recovery. And that elected node will read the redo logs of the crashed instance and do the recovery.

However in 19c,  One instance will recovery buddy of another instance. like.

Instance A is recovery buddy of instance B.

Instance B is recovery buddy of instance C.

Instance C is recovery buddy of instance A.

And this buddy instance will the track the block/redo changes of the mapped instance and keep them in its sga( in hash table ).

So recovery buddy features helps in reducing the recovery time( as it eliminates the elect and redo read phase).

**89. What is nodeapps?**

Nodeapps are standard set of oracle application services which are started automatically for RAC.

Node apps Include: vip,network,adminhelper,ONS

**90. CSSD is not coming up ? What you will check and where you will check.**

1. Voting disk is not accessible
2. Issue with private interconnect

2.the auto\_start parameter is set to NEVER in ora.ocssd resource . ( To fix the issue, change it to always using crsctl modify resource )

**91. How you check the cluster status?**

crsctl stat res -t

crsctl check crs

crsctl stat res -t -init

**93. What are the different types of heart beats in Oracle RAC?**

There are two types of heart beat.

**Network heartbeat** is across the interconnect, every one second, a thread (sending) of CSSD sends a network tcp heartbeat to itself and all other nodes, another thread (receiving) of CSSD receives the heartbeat. If the network packets are dropped or has error, the error correction mechanism on tcp would re-transmit the package, Oracle does not re-transmit in this case. In the CSSD log, you will see a WARNING message about missing of heartbeat if a node does not receive a heartbeat from another node for 15 seconds (50% of misscount). Another warning is reported in CSSD log if the same node is missing for 22 seconds (75% of misscount) and similarly at 90% of misscount and when the heartbeat is missing for a period of 100% of the misscount (i.e. 30 seconds by default), the node is evicted.

**Disk heartbeat** is between the cluster nodes and the voting disk. CSSD process in each RAC node maintains a heart beat in a block of size 1 OS block in a specific offset by read/write system calls (pread/pwrite), in the voting disk. In addition to maintaining its own disk block, CSSD processes also monitors the disk blocks maintained by the CSSD processes running in other cluster nodes. The written block has a header area with the node name and a counter which is incremented with every next beat (pwrite) from the other nodes. Disk heart beat is maintained in the voting disk by the CSSD processes and If a node has not written a disk heartbeat within the I/O timeout, the node is declared dead. Nodes that are of an unknown state, i.e. cannot be definitively said to be dead, and are not in the group of nodes designated to survive, are evicted, i.e. the node’s kill block is updated to indicate that it has been evicted.

Reference – <https://databaseinternalmechanism.com/oracle-rac/network-disk-heartbeats/>

[grid@Linux-01 ~]$ crsctl get css misscount  
CRS-4678: Successful get misscount 30 for Cluster Synchronization Services.

[grid@Linux-01 ~]$ crsctl get css disktimeout  
CRS-4678: Successful get disktimeout 200 for Cluster Synchronization Services.

**4. Explain how client connection is established in  RAC database ?**

LREG process on each instances registers the database service of the node with default local listener and scan listener. The listeners store the workload information of each node.

So when client tries to connect using scan\_name and port,

1. scan\_name will be resolved through DNS , which will redirect to 1st scan ip ( out of 3).
2. the client will connected to the respective scan listener
3. The scan listener compares the work load of both the instances and if scan determines that node1 has least load , then scan listener send the vip address  and port details of that particular nodes local listener to client.
4. Now client connects to that local listeners and a dedicated server process is created
5. Client connection becomes successful and it starts accessing the database.