ORACLE11gRAC

# 1,集群体系结构

## 1.1 集群简介

一个并行集群系统，整个系统由Oracle Clusterware（集群就绪软件）和Real Application Clusters（RAC）两部分组成。

本质：位于不同操作系统的oracle实例节点同时访问同一个数据库，每个节点间通过私有网络进行通信，互相监控节点的运行状态，oracle所有的数据文件，联机日志文件，控制文件等均放在集群的共享存储设备上，共享存储设备可以是RAW裸设备,ASM,OCFS2集群文件系统,NAS网络区域存储,NFS等，所有集群节点可以同时读写共享存储，

Oracle10g起，oracle提供自己的集群软件，oracle clusterware，通过CRS（cluster Ready Services）

### 1.1.1 单实例与RAC环境对比

RAC环境中每个实例有其自己的系统全局区SGA和后台进程，但是所有数据文件和控制文件都可供所有节点同等访问，所以必须把这些文件放在一个共享磁盘子系统上，每个实例还有其自己的专用联机重做日志文件，这些联机重做日志文件只能由它所属实例写入，但是在实例故障恢复期间， 这些文件必须可以供其他实例读取，所以需要驻留在一个共享磁盘子系统上，不是位于节点本地存储上。

|  |  |  |
| --- | --- | --- |
| 组件 | 单实例 | RAC环境 |
| SGA | 实例自己的SGA | 每个实例有SGA |
| 后台进程 | 自己的后台进程集 | 每个实例有后台进程集 |
| 数据文件 | 一个实例本地访问 | 共享存储 |
| 控制文件 | 一个实例本地访问 | 共享存储 |
| 联机重做日志 | 单实例写入和读取 | 只有一个实例可写，其他实例可以在恢复和存档期间读取，如果一个实例关闭，其他实例的日志切换可以强制对空闲实例重做日志存档 |
| 存档后的重做日志 | 单实例使用 | 专属该实例，但在媒介恢复期间其他实例需要访问所需的存档日志 |
| 快闪恢复日志 | 单实例访问 | 所有实例共享，共享存储中 |
| 告警日志和跟踪文件 | 单实例使用 | 专属每个实例，其他实例永远不会读写 |
| Oracle\_home | 同一台计算机上访问不同库的多个实例可使用相同的执行文件 | 与单实例相同，可以放在共享文件系统上，允许RAC中所有使用共用 |
|  |  |  |

## 1.2 Clusterware简介

CRS主要完成集群成员，心跳监控，故障切换等功能，CRS要求每个集群节点的操作系统必须相同。

CRS的主要两个集群组件：voting disk 和 Oracle Cluster Registy

### 1.2.1 voting disk

即为表决磁盘，集群中每个节点定期评估自身的健康状态，然后会把它的状态信息放入到表决磁盘上，并且节点间也会互相查看运行状态，并把信息传递给其他节点而进入表决磁盘。当集群发生故障时，通过表决磁盘进行投票仲裁等，因此表决磁盘必须放在共享存储上。表决磁盘可以是一个裸磁盘分区，也可以是独立文件，一般大小在10-20m左右即可

### 1.2.2 Oracle Cluster Registy

OCR，简称集群注册服务，OCR主要用于记录RAC中集群和数据库的配置信息，包括集群节点的列表，集群数据库实例到节点的映射以及CRS应用程序资源信息，CRS使用两种心跳设备来验证成员状态。

一个是表决磁盘，集群同步服务进程每隔几秒都会向表决磁盘写入一条心跳信息，集群通过表决磁盘访问验证节点状态，如果某个节点没有向表决磁盘写入信息，集群认为节点失效了。

另一个是心跳节点私有以太网的心跳，通过心跳判断节点是否出现网络故障。

CRS建议用于内部通讯的私有以太网心跳必须与用于RAC节点间通讯的网络分开，不能在同一个网络中。

### 1.2.3 Clusterware进程介绍

#### 1，Cluster Synchronization Services

CSS，用于管理与协调集群中各节点的关系，并用于节点通信，当节点在加入或离开集群时，由CSS进行通知集群，对应的后台进程为CSSD，该进程由oracle用户运行和管理，当节点发生故障，CSSD会自动重启操作系统。

OCSSD提供节点之间同步服务，其故障会导致计算机重新启动，以oracle用户身份运行，以两种心跳机制提供服务，网络心跳和磁盘心跳

#### 2，Cluster Ready Services

CRS，是管理群内高可用操作的主要程序，在集群中CRS管理所有资源，包括数据库，服务，实例，VIP地址，监听，应用进程等，CRS在集群中对应的后台进程是CRSD，该进程可以对集群资源进行启动，停止，监视和容错等操作，正常下，CRSD监控节点各种资源，异常时，自动重启或切换该资源。

CRSD以root用户身份运行，为每个实例生成一个专用的RACGIMON进程，以监控数据库和ASM实例

CRS需要公共接口，私有接口，和虚拟IP才能运行，互相ping

#### 3，Process Monitor Daemon

OPROCD，此进程被锁定在内存中，作为一个实时进程运行，用于监控集群及提供I/O防护，进程运行在每个节点上，定期健康检测，如果在超过希望的间隔内，仍然不能通信，则进程会重置处理器及重启节点，进程故障将导致Clusterware重启节点，以root用户运行，该进程故障会导致节点重启

OPROCd进程很重要，集群件使用另一个OCLSOMON进程来监控，若OPROCd挂起，就导致集群节点重启

#### 4，Oracle Notification Service

ONS，通告服务，用于发布和订阅Fast Application Notification事件。

在安装集群件期间配置，CRS启动时在每个集群节点上启动该进程，当集群资源状态发生变化，每个集群上的ONS进程会互相通信，交换高可用性事件信息FAN

#### 5，Event Management

EVM ,是一个事件检测的后台进程，由Oracle运行和管理。

EVMD的守护进程生成一个evmlogger永久子进程，在出现状况时生成事件，根据需要生成新的子进程，进程发生故障时会重新启动，evmd终止不会使实例停止，oracle用户身份运行。

|  |  |  |  |
| --- | --- | --- | --- |
| CRS进程 | 功能 | 进程故障时 | 角色 |
| 守护进程CRSd | 资源监控，资源故障转移和节点恢复 | Reboot模式下会重新启动，restart模式下不会导致节点重启 | root |
| 守护进程CSSd | 基本节点成员资格，组服务和基本锁定 | 节点重启 | Oracle |
| ONS | 用于向客户端扩展高可用性通知 | 故障期间自动重启 | Oracle |
| 守护进程EVMd | 生成一个子进程事件记录器，并生成标注 | 自动重启，不会导致节点重启 | Oracle |
| 守护进程OPROCd | 提供基本的集群完整性服务 | 节点重启 | root |

## 1.3 RAC数据库体系结构

### 1.3.1 RAC简介

Rac是一个具有共享缓存体系结构的集群数据库，与Clusterware或第三方集群软件共同组成Oracle集群系统。

它是一个全共享式的体系架构，所有的数据文件，控制文件，联机日志文件，参数文件等都必须存放在共享磁盘中，让所有节点可以访问到，RAC支持多种存储方式，可以使用如下任意一种：

1. 裸设备(Raw)

不经过文件系统，读写性能高，数据维护和备份不方便

1. 集群文件系统(OCFS2)

Oracle开发的集群文件系统OCFS，用于windows,Linux和Solaris，已发展到ocfs2

1. 网络文件系统NFS
2. ASM，Automated Storage Management

是oracle推荐的共享存储方式，ASM其实就是RAW方式存储数据，加入了数据管理功能，将数据直接写入磁盘，避免经过文件系统产生IO，可以方便的管理共享，提供异步IO性能，ASM可以通过分配IO负载来优化性能，避免手动调整IO。

### 1.3.2 RAC的特点

1，实现多节点间负载均衡

2，提供高可用服务

3，通过横向扩展提供了并发连接数

4，通过并行执行技术提供了事务响应时间

1. 扩展性好

### 1.3.3 RAC的缺点

1，维护复杂，维护人员要求高

2，底层规划设计不好，整体性能会差

3，集群多个节点，需要多台服务器，增加了成本

### 1.3.4 RAC的后台进程

RAC每个节点就是一个数据库实例，每个实例都有自己的后台进程和内存结构，并且相同，也有不同之处：

1. 每个实例至少拥有一个额外的重做线程redo thread
2. 每个实例都有自己的撤销表空间undo tablespace

其实每个实例的SGA内有一个buffer cache，通过cache Fusion缓存融合技术，RAC在各个节点之间同步SGA中的缓存信息，保证数据一致性。

1. 数据共享进程GCS 和 GES

Global cache Service 全局缓存服务和 Global enqueuer Service全局队列服务，他们是最基本的RAC进程，主要用于协调对共享数据库和数据库内的共享资源的同时访问，它们使用全局资源目录GRD,Global Resource Directory 来记录和维护每个数据文件的状态信息，而GRD保存在内存中，内容分布存储在所有实例上，每个实例管理部分内容

全局缓存服务GCS，一个节点的缓冲区缓存中可能包含了经常被另一个节点请求的数据，数据共享与交互的管理工作由它完成

全局资源目录GRD，集群组中所有资源构成一个集中资源仓库，所有实例加起来构成GRD,它由2个服务管理：

全局缓存服务GCS，和全局队列服务GES，RAC中几个特别的进程和GRD结合，使得RAC可以使用缓存融合技术，进程是：

LMS 全局缓存服务进程

LMON 全局队列服务监控器

LMD 全局队列服务守护进程

LCK0 实例队列进程

DIAG 诊断守护进程

Lock Manager Server(LMSn)

是缓存融合中使用的一个进程，在特定情况下，它可以从数据库所在实例的缓冲区向请求实例的缓冲区缓存中传递数据块的一致性副本，不需要进行磁盘写入，还可以从LMD建立的服务器队列中获取请求，以执行所请求的锁操作，管理对GCS资源的锁管理器服务器请求，将请求发送到一个LMS进程处理的服务队列，负责全局锁的死锁检测，并监控锁对话的超时

Lock Monitor(LMON)

锁监控器进程，负责管理全局队列服务GES，并负责在有实例加入或离开时进行集群重新配置和锁重新配置，检查实例死亡并监听本地消息，生成详细的跟踪文件，用于跟踪实例重新配置，后台LMON进程监控整个集群，管理全局资源

Lock Manager Daemon(LMD)

管理对全局队列和全局资源的访问，并更新相应的队列状态，处理来自其他实例的资源请求

是GCS管理队列管理服务请求的守护进程，资源代理进程管理对资源的请求，以控制对数据块的访问，LMD进程负责死锁检测和来自另外一个实例的远程资源请求。

Lock Processes(LCK)

管理实例间资源请求和跨实例调用操作，管理除cache fusion以外的资源请求，比如library和row chae

一个实例只能使用一个LCK进程，因为主要功能由LMS进程处理

Diagnosability Daemon(DIAG)

捕获实例中失败进程的诊断信息，并生成相应的TRACE文件。

### 1.3.5 RAC存储规划

涉及软件有Oracle Clusterware , RAC数据库软件，涉及voting disk,OCR等，磁盘空间大小如下：

|  |  |  |
| --- | --- | --- |
| 软件名称 | 用途 | 磁盘空间 |
| Clusterware软件 | 集群管理软件 | 800M |
| RAC数据库软件 | RAC运行程序 | 2-3G |
| RAC数据库实例 | RAC共享数据库文件 | 5-8G |
| Voting Disk表决磁盘 | 记录集群节点信息 | 20-50M |
| OCR集群注册服务 | 存储集群配置信息 | 100-200M |
| Flash Recovery Area | 快速恢复数据 | 2G |

推荐三种常用的存储方式：

|  |  |  |
| --- | --- | --- |
| 软件名称 | 存储类型 | 位置 |
| Clusterware软件 | Ext3 | 本地 |
| RAC数据库软件 | Ext3 | 本地 |
| RAC数据库实例 | ASM | 共享磁盘 |
| Voting Disk表决磁盘 | RAW | 共享磁盘 |
| OCR集群注册服务 | RAW | 共享磁盘 |
| Flash Recovery Area | ASM | 共享磁盘 |

|  |  |  |
| --- | --- | --- |
| 软件名称 | 存储类型 | 位置 |
| Clusterware软件 | Ext3 | 本地 |
| RAC数据库软件 | OCFS2 | 共享磁盘 |
| RAC数据库实例 | ASM | 共享磁盘 |
| Voting Disk表决磁盘 | OCFS2 | 共享磁盘 |
| OCR集群注册服务 | OCFS2 | 共享磁盘 |
| Flash Recovery Area | ASM | 共享磁盘 |

|  |  |  |
| --- | --- | --- |
| 软件名称 | 存储类型 | 位置 |
| Clusterware软件 | Ext3 | 本地 |
| RAC数据库软件 | OCFS2 | 共享磁盘 |
| RAC数据库实例 | OCFS2 | 共享磁盘 |
| Voting Disk表决磁盘 | OCFS2 | 共享磁盘 |
| OCR集群注册服务 | OCFS2 | 共享磁盘 |
| Flash Recovery Area | OCFS2 | 共享磁盘 |

## 1.4 RAC安装

### 1.4.0 注意步骤事项

ASMLIB包（红帽7不再支持）：

oracleasmlib-2.0.4-1.el5.i386.prm

oracleasm-2.6.18-194.11.1.el5-2.0.5-1.el5.i686.rpm

oracleasm-support-2.1.3-1.el5.i386.rpm

主机解析文件

/etc/hosts中配置主机IP解析，域名解析

必备软件包

系统内核参数

若存在NTP服务器，需要在集群节点上配置NTP配置文件（财务管控一体机）

# vim /etc/ntp.conf 加入本地NTP服务器地址

server 10.137.252.67 perfer

# ntpq -p 查询网络中的NTP服务器，同时显示客户端和每个服务器的关系

# systemctl restart ntpd

# yum install ntp ntpdate -y

# ntpdate -u ntp.zpepc.com.cn

# systemctl status ntpdate

# systemctl status ntpd

### 1.4.1 主机方面

CentOS 7+RAC11g. VM虚拟机

修改主机名

## 1.5 CRS的管理和维护

常用命令crs\_stat,crs\_start,crs\_stop,crsctl等。

### 1.5.1 crs\_stat命令

1，查看集群状态

crs\_stat -t 概要资源信息

crs\_stat -l | head -n 15 详细资源信息

2，关闭服务资源

crs\_stop RESOURCE 停止某个资源

crs\_start RESOURCE 启动某个资源

如：

crs\_stop ora.rac1.LISTENER\_RAC1.lsnr 停止RAC1上的监听

3, 关闭集群所有资源(时间比较长)

crs\_stop -all

crs\_start -all

### 1.5.2 crsctl命令

功能强大的命令，可以检查CRS后台进程运行状态，添加/删除表决磁盘，启动关闭集群所有资源，启动关闭CRS服务等。

crsctl check crs (css, evm, ctss) 检查服务状态

crsctl check cluster [[-all]|[-n <server>[...]]] 检查集群所有资源状态

crsctl query crs activeversion 检查CRS运行版本

crsctl lsmodules css 查看css加载的模块

crsctl stop resources 停止本节点所有CRS资源

crsctl stop crs 关闭CRS服务，需要root

crsctl query css votedisk 查看表决磁盘的信息（

备份：dd if= /dev/asm-diskb1 of=/tmp/votedisk.bak

恢复： dd if=/tmp/votedisk.bak of=/dev/asm-diskb1

删除： crsctl delete css votedisk /dev/asm-diskb1

添加： crsctl add css votedisk /dev/asm-diskb1）

### 1.5.3 ocr检查

ocrcheck

ocrconfig –showbackup 查看ocr备份

ocrconifg –restore /u01/app/11.2.0/grid/cdata/rac-cluster/backup00.ocr 恢复ocr

ocr文件信息：/etc/oracle/ocr.loc

## 1.6 ASM管理

### 1.6.1 特点

1，自动调整IO负载：

ASM可以在所有可用的磁盘中自动调整IO负载，避免人工调整的难度，优化了性能，利用ASM可以在线增加数据库大小，无需关闭数据库

2，条带化存储：

ASM将文件分为多个分配单元，简称AU进行存储，并在所有磁盘间平均分配每个文件的AU

3，在线自动负载均衡：

当共享存储设备有变化时，ASM中的数据会自动的均匀分配到现有存储设备中，同时还可以调节数据的负载均衡速度

4，自动管理数据库文件：

在ASM存储管理中，oracle数据文件是ASM自动管理的，ASM创建的任何文件一旦不再需要，就会自动删除，但是ASM不管理二进制文件，跟踪文件，预警日志和口令文件

5，数据冗余：

ASM通过磁盘组镜像可以实现数据冗余，而不需要第三方工具来实现

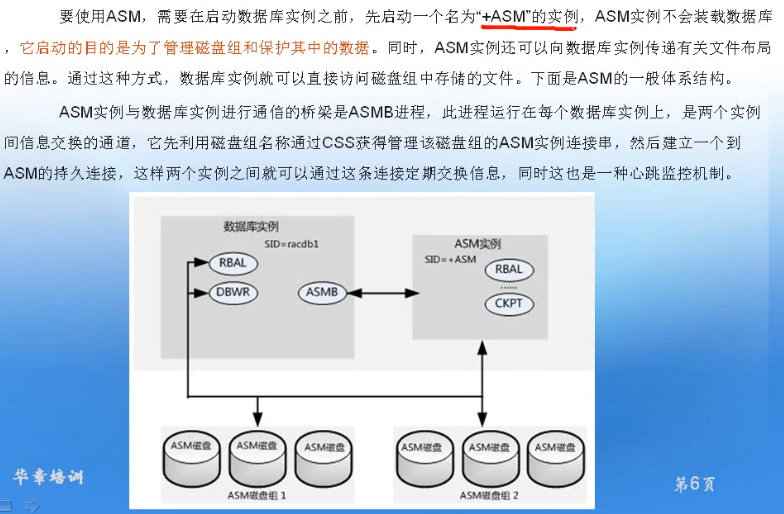
6，支持各种oracle数据文件：

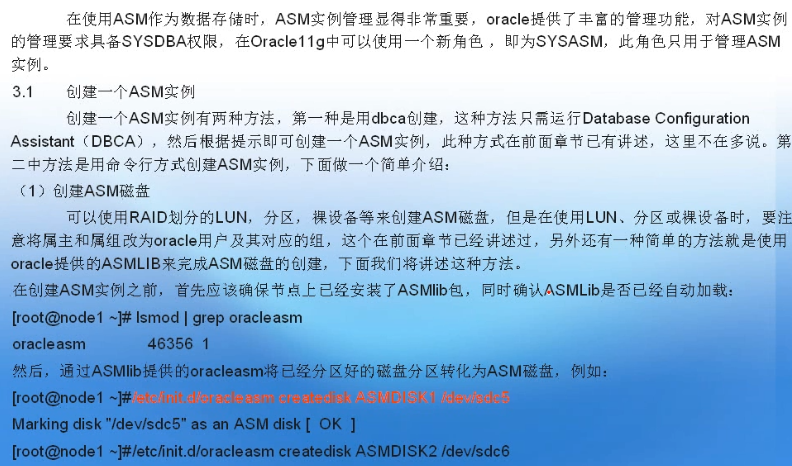
ASM存储支持oracle数据文件，日志文件，控制文件，归档文件，RMAN备份集等

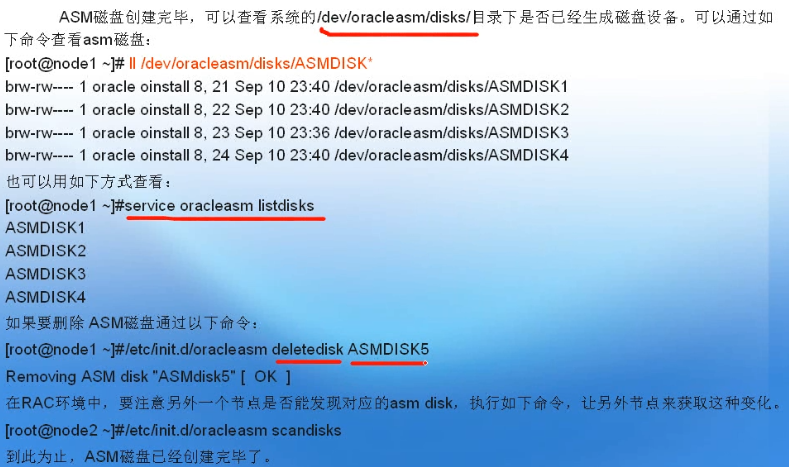
### 1.6.2 ASM体系结构

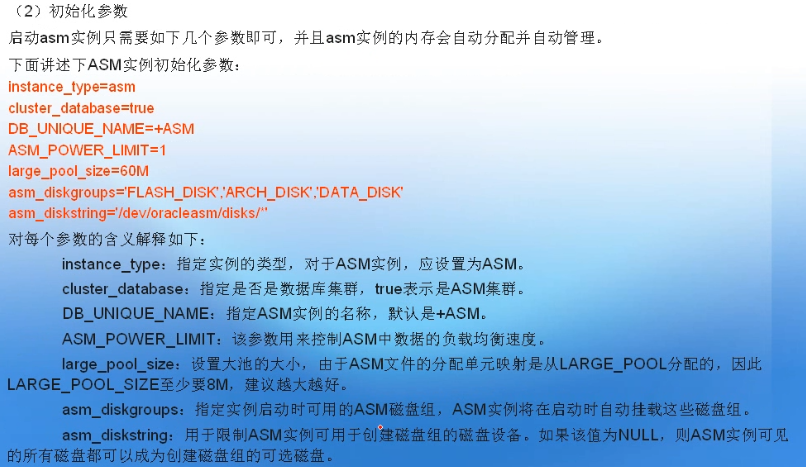


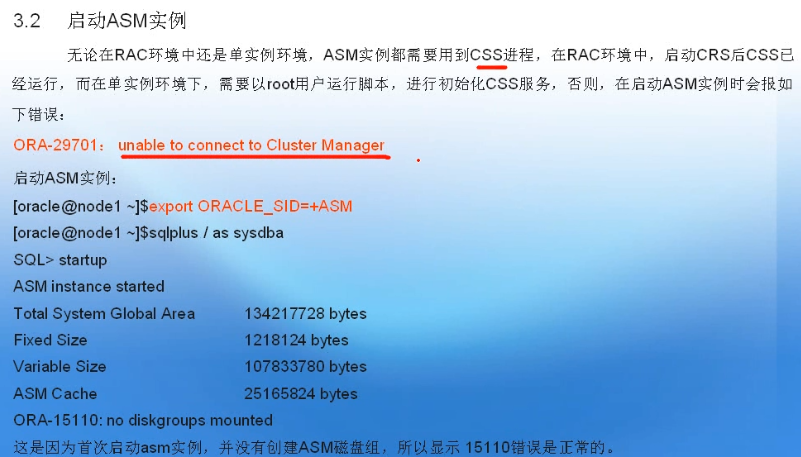
最顶层是ASM磁盘组，ASM实例和数据库实例可以直接访问这些磁盘组，然后是ASM文件，每个ASM文件只能包含在一个磁盘组中，不过，一个磁盘组中可以包含属于多个数据库的多个ASM文件，并且单个数据库可以使用来自多个磁盘组的存储空间，第三部分是ASM磁盘，多个ASM磁盘组成ASM磁盘组，但每个ASM磁盘只能属于一个磁盘组，分配单元AU，分配单元是ASM磁盘组分配的最小连续磁盘空间，ASM磁盘是按照AU进行分区，每个AU的大小为1MB，在这个结构的最底层是oracle数据块，由于AU是ASM分配的最小的连续磁盘空间，因此ASM是不允许跨分配单元拆分一个oracle数据块的。



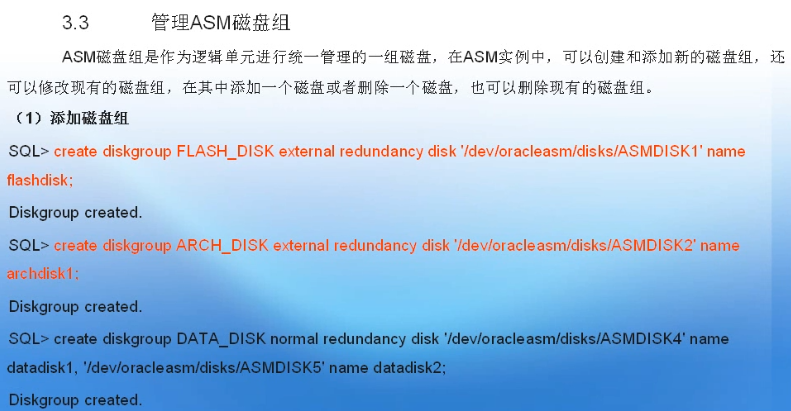


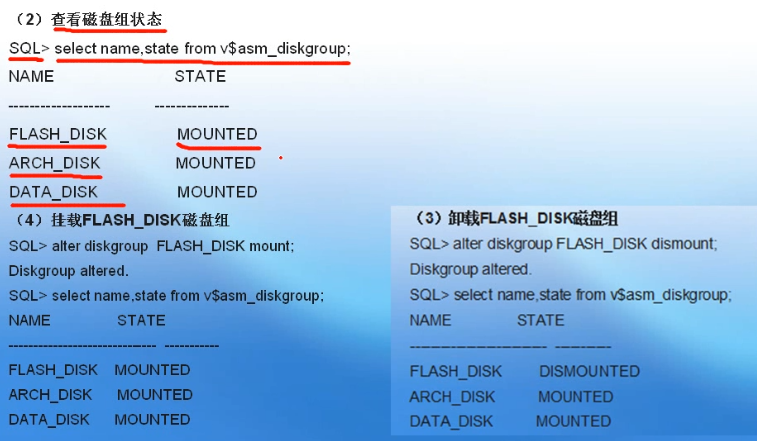


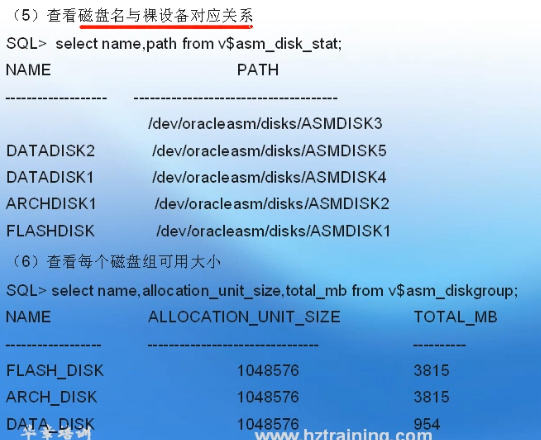


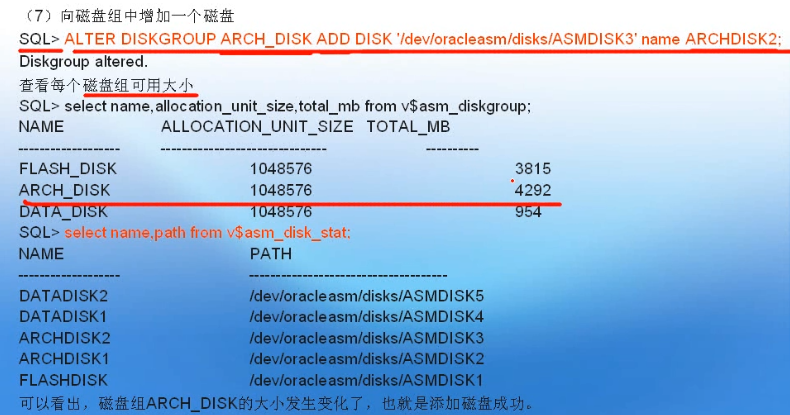


### 1.6.3 管理ASM磁盘组









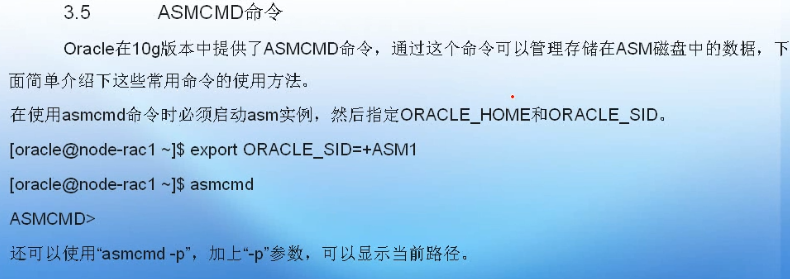




#### 1.6.3.1 ASM镜像选项

|  |  |  |
| --- | --- | --- |
| ASM冗余 | 镜像级别 | 描述 |
| 外部 | 无镜像 | 没有ASM镜像，可以在存储级别或磁盘级别实现镜像时使用 |
| 常规 | 双重镜像 | 每个盘区在不同的故障群组中有一个镜像副本 |
| 高 | 三重镜像 | 每个盘区在不同的故障群组中有两个镜像副本 |

### 1.6.4 ASMCMD命令



### 1.6.5 创建ASM磁盘

#### 一．Asmlib驱动

在EL7只需要安装2个包，安装之后出现oracleasm命令

1，oracleasm-support-2.1.11-2.el7.x86\_64.rpm，

2，oracleasmlib-2.0.12-1.el7.x86\_64.rpm

#### 二，Asmlib驱动配置

[root@rac1 ~]# oracleasm configure -i

Default user to own the driver interface []: grid --磁盘设备默认用户

Default group to own the driver interface []: asmadmin --磁盘设备默认组

Start Oracle ASM library driver on boot (y/n) [n]: y --开机启动ASM库

Scan for Oracle ASM disks on boot (y/n) [y]: y --开机挂在ASM磁盘

Writing Oracle ASM library driver configuration: done

[root@rac1 ~]# oracleasm configure 查看配置

[root@rac1 ~]# oracleasm init 初始化配置

[root@rac1 ~]# lsmod | grep oracleasm 查看加载模块

#### 三，ASM磁盘和磁盘组

1，对未使用的分区进行创建磁盘(tail -f /var/log/oracleasm)

# oracleasm createdisk ASMDISK1 /dev/sdf5

# oracleasm createdisk ASMDISK2 /dev/sdf6

# service oracleasm listdisks 查看ASM磁盘

# oracleasm scandisks 获取其他节点磁盘信息

2，删除ASM磁盘

# oracleasm deletedisk ASMDISK1

3, 创建磁盘组

# create diskgroup FLASH\_DISK external redundancy disk ‘/dev/oracleasm/disks/ASMDISK1’ name flashdisk; 多个磁盘在最后加入磁盘信息(’’name ..),用逗号分隔

4, 挂载和卸载ASM磁盘组

# alter diskgroup DATA\_DISK dismount;

# alter diskgroup DATA\_DISK mount; 慢一点

5，磁盘组中增加磁盘

# alter diskgroup ARCH\_DISK add disk ‘/dev/oracleasm/disks/archdisk3’ name ARCHDISK2;

6, 磁盘组中删除磁盘

# alter diskgroup ARCH\_DISK drop disk ARCHDISK2;

7, 删除磁盘组

# drop diskgroup FLASH\_DISK;

## 1.7 RAC集群功能

### 1.7.1 负载均衡测试

RAC数据库的负载均衡是指对数据库连接的负载均衡，分为两种：1，基于客户端连接的负载均衡，2，服务端的负载均衡。

#### 客户端负载均衡

该方式与实例的负载和监听没有关系，集群节点不需要任何设置，需要在客户端机器上tnsnames.ora文件中添加负载均衡策略配置。在配置RAC负载均衡时，客户端连接的是RAC的服务名，而不是实例名，

Select global\_name from global\_name; 服务端获取RAC的服务名

客户端配置的tnsnames.ora文件如下：



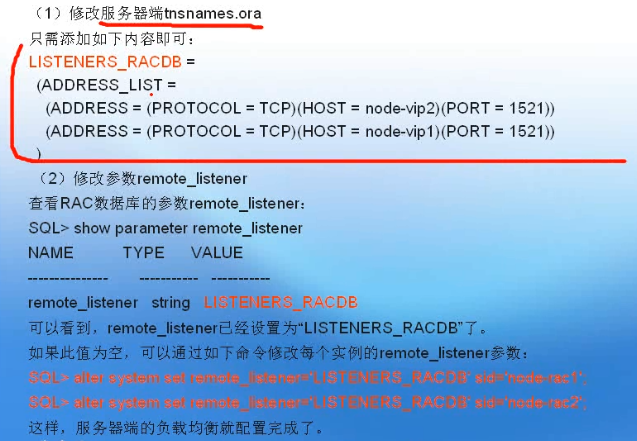
LOAD\_BALANCE=yes 启用连接负载均衡，

#### 服务端负载均衡

客户端负载实现后，由于客户端机器不了解RAC负载情况，可能会连接至负荷较大的节点，并增加连接数。

服务端负载情况由PMON进程来定期更新，PMON每3秒会将集群每个节点的负载信息及连接数写入service\_register中，当节点负载发生变化时，立刻通知监听程序，由监听决定新的连接分配到哪个接待你，最后实现真正的负载均衡。

服务端配置需要在tnsnames.ora里添加一个连接到各个节点的监听配置，然后初始化参数中设置remote\_listeners，两个节点同步配置信息



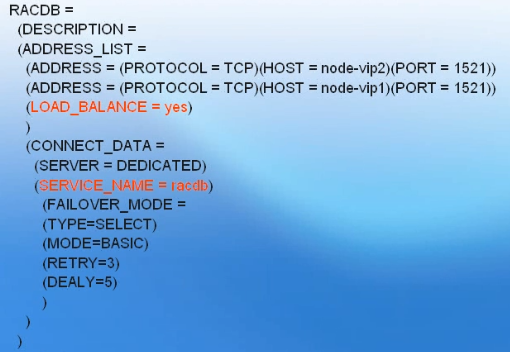
### 1.7.2 透明应用失败切换

简称TAF(Transparent Application Failover)，是客户端的一种功能。TAF包含两层意思：

失败切换：是指客户端连接到某个实例，如果连接失败，可以连接到另外一个实例

透明应用：是指客户端应用程序在连接失败后可以自动重新连接到另外一个数据库实例，而这个过程对应用程序不可见

要使用TAF功能，只需要修改客户端中tnsnames.ora文件的配置即可，参考设置如下



参数解释：

TYPE：这个是指定FAILOVER\_mode的类型，有三种session,select和none

1. Session 表示当一个正在连接的会话实例发生故障时，系统可以自动将会话切换到其他可用的实例，而应用程序无须再次发起连接请求，但是实例故障时正在执行的sql需要重新执行
2. Select 表示如果正在连接的实例发生故障，将使用游标和之前的快照继续执行select操作，其他操作需要重新执行
3. None 这个是客户端默认值，表示禁止sql接管功能

MODE：表示连接模式，有两种类型，分别是basic和preconnect

1. Basic表示在建立初始连接时仅连接到一个节点，并且只有在发生节点故障时才连接到备用节点
2. Preconnect表示在建立初始连接时就连接到主节点和备用节点

RETRY：表示当前节点失败后，失败切换功能尝试连接备用节点的次数

DELAY：表示两次尝试之间的等待秒数