**Oracle Clusterware**

云和恩墨(北京)信息技术有限公司

技术顾问 燕鑫

http://www.enmotech.com

**文档控制：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序** | **版本号** | **更改人** | **日期** | **备注** |
| 1 | 1.0版 | 燕鑫 | 2018-11-1 | 初始版本 |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[1. 集群概要 - 6 -](#_Toc530667232)

[1.1 Oracle Clusterware - 6 -](#_Toc530667233)

[1.1.1 Oracle Flex Clusters - 6 -](#_Toc530667234)

[1.1.2 集群的优势 - 7 -](#_Toc530667235)

[1.1.3 Deployment Architectures - 7 -](#_Toc530667236)

[1.2 System Requirements - 9 -](#_Toc530667237)

[1.2.1 硬件要求 - 9 -](#_Toc530667238)

[1.2.2 操作系统要求 - 9 -](#_Toc530667239)

[1.2.3 软件要求 - 9 -](#_Toc530667240)

[1.2.4 网络配置要求 - 10 -](#_Toc530667241)

[1.3 Platform-Specific Software Components - 11 -](#_Toc530667242)

[1.3.1 Technology Stack - 11 -](#_Toc530667243)

[1.3.2 集群启动流程 - 13 -](#_Toc530667244)

[1.4 安装Oracle Clusterware概要 - 15 -](#_Toc530667245)

[1.5 升级和打补丁概要 - 15 -](#_Toc530667246)

[1.6 Grid Infrastructure Management Repository - 15 -](#_Toc530667247)

[1.7 Domain Services Clusters - 17 -](#_Toc530667248)

[1.8 管理集群环境概要 - 17 -](#_Toc530667249)

[1.9 Command Evaluation - 18 -](#_Toc530667250)

[1.10 Grid环境中Cloning and Extending Oracle Clusterware - 19 -](#_Toc530667251)

[1.11 Cluster Time Management - 19 -](#_Toc530667252)

[1.11.1 概述 - 19 -](#_Toc530667253)

[1.11.2 step time synchronization - 20 -](#_Toc530667254)

[1.11.3 slew time synchronization - 20 -](#_Toc530667255)

[1.11.4 Active and Deactive CTSS - 20 -](#_Toc530667256)

[2. 集群的配置和管理 - 21 -](#_Toc530667257)

[2.1 Role-Separated Management - 21 -](#_Toc530667258)

[2.1.1 Managing Cluster Administrators - 22 -](#_Toc530667259)

[2.1.2 Configuring Role Separation - 22 -](#_Toc530667260)

[2.2 使用Grid Setup Wizard配置GI - 22 -](#_Toc530667261)

[2.2.1 Configuring a Single Node - 23 -](#_Toc530667262)

[2.2.2 Configuring Multiple Nodes - 23 -](#_Toc530667263)

[2.2.3 Upgrading Oracle Grid Infrastructure - 23 -](#_Toc530667264)

[2.2.4 Running the Configuration Wizard in Silent Mode - 23 -](#_Toc530667265)

[2.3 Server Weight-Based Node Eviction - 23 -](#_Toc530667266)

[2.4 Oracle Database QoS Management - 24 -](#_Toc530667267)

[2.5 Grid Naming Service（GNS） - 24 -](#_Toc530667268)

[2.6 Node Failure Isolation - 24 -](#_Toc530667269)

[2.7 理解手动配置网络中的网络地址 - 24 -](#_Toc530667270)

[2.7.1 理解网路地址配置要求 - 24 -](#_Toc530667271)

[2.7.2 理解SCAN地址和Client Service Connections - 25 -](#_Toc530667272)

[2.7.3 SCAN listeners和使用valid node checking限制服务注册 - 26 -](#_Toc530667273)

[2.8 手动改变IP地址 - 26 -](#_Toc530667274)

[2.8.1 用SRVCTL改变VIP - 26 -](#_Toc530667275)

[2.8.2 修改私网配置（oifcfg） - 27 -](#_Toc530667276)

[2.8.3 用SRVCTL创建一个网络 - 28 -](#_Toc530667277)

[2.8.4 集群中的网络配置 - 28 -](#_Toc530667278)

[2.8.5 IPv4和IPv6的转换 - 28 -](#_Toc530667279)

[3. Policy-Based Cluster - 29 -](#_Toc530667280)

[4. Oracle Flex Clusters - 29 -](#_Toc530667281)

[5. Rapid Home Provisioning, Scaling, Patching, and Upgrading - 29 -](#_Toc530667282)

[6. 管理OCR和Voting Files - 29 -](#_Toc530667283)

[6.1 管理OCR和OLR - 29 -](#_Toc530667284)

[6.1.1 OCR - 31 -](#_Toc530667285)

[6.1.2 OLR - 37 -](#_Toc530667286)

[6.2 管理voting files - 39 -](#_Toc530667287)

[6.2.1 Voting Files的存储 - 39 -](#_Toc530667288)

[6.2.2 三个不能 - 40 -](#_Toc530667289)

[6.2.3 备份voting files - 40 -](#_Toc530667290)

[6.2.4 恢复voting files - 40 -](#_Toc530667291)

[7. 增删集群节点 - 41 -](#_Toc530667292)

[8. 克隆Oracle Clusterware - 41 -](#_Toc530667293)

[9. Making Applications Highly Available Using Oracle Clusterware - 41 -](#_Toc530667294)

[9.1 Oracle Clusterware Resources and Agents - 41 -](#_Toc530667295)

[9.1.1 Oracle Clusterware Resources - 42 -](#_Toc530667296)

[9.1.2 Oracle Clusterware Resource Types - 42 -](#_Toc530667297)

[9.1.3 Agents in Oracle Clusterware - 42 -](#_Toc530667298)

[9.1.4 Action Scripts - 42 -](#_Toc530667299)

[9.1.5 Building an Agent - 42 -](#_Toc530667300)

[9.1.6 Registering a Resource in Oracle Clusterware - 42 -](#_Toc530667301)

[9.2 Overview of Using Oracle Clusterware to Enable High Availability - 42 -](#_Toc530667302)

# 集群概要

## Oracle Clusterware

Oracle Clusterware是一款能提供综合的多层的高可用，并能对所有集群下的环境的资源进行统一管理的软件。它可以把很多独立的服务器联合到一起对外提供服务。

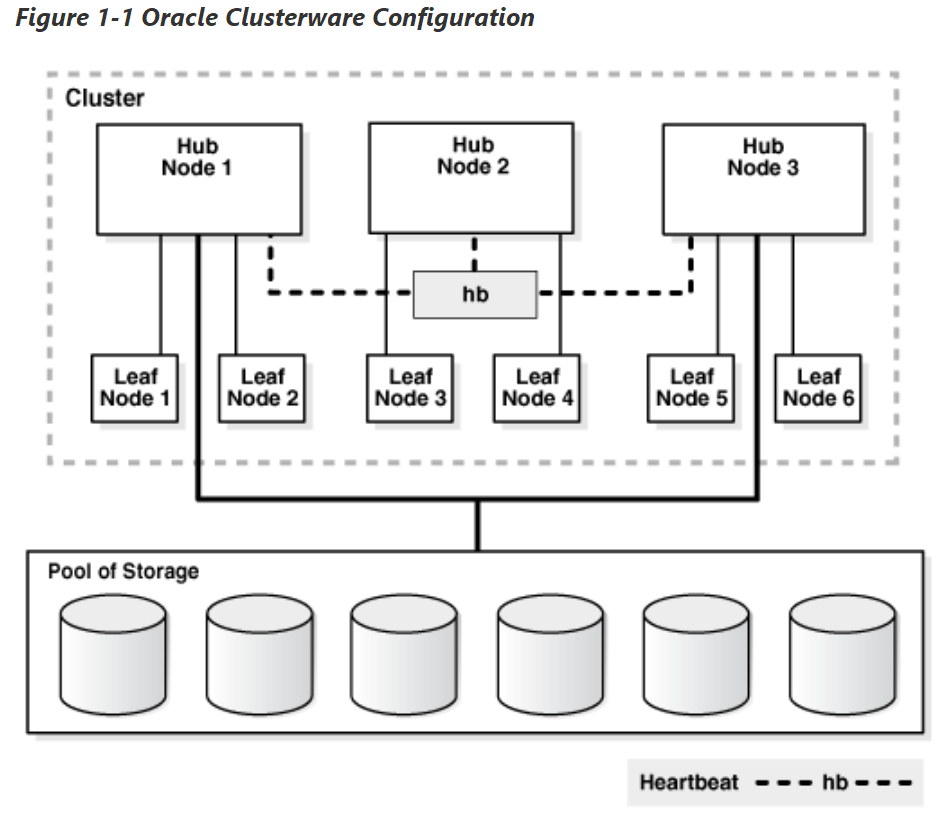
Oracle Clusterware是RAC的基础结构，同时也是所有集群下的平台上的应用的资源管理和高可用架构。

Oracle Clusterware是10.1引进的，目的就是实现多实例访问数据库（RAC），同时在云端提供企业级别的弹性要求和动态的、在线的计算资源分配。

### Oracle Flex Clusters

从12.2开始，所有的集群都是flex集群，老版本的集群在升级到12.2以后，也会转化成flex集群，这也就意味着，12.2以后，ASM也都是Oracle Flex ASM（Flex Cluster的必要条件）。同时也代表12.2的集群的节点分为hub nodes和leaf nodes。

如下是Oracle Clusterware的架构：



我们可以看到，Oracle Flex Clusters包含有俩种类型的节点：hub节点和leaf节点（a hub and spoke architecture）。Oracle Flex Cluster最多可以有64个hub nodes，每个hub nodes可以有很多leaf nodes。Hub nodes和leaf nodes可以用来支撑不通的应用。

Hub nodes和共享存储进行直接访问，而leaf nodes要通过hub nodes才可以同共享存储通信。

集群中hub nodes是必需的，leaf nodes是可以选的，如果想要有leaf nodes，则至少得有一个hub nodes。

### 集群的优势

集群的优势包括：

1. 应用们的伸缩性，比如通过服务控制应用对节点的访问
2. 支持故障转移
3. 提高生产力
4. 能够控制应用的启动顺序，保证依赖的进程按顺序启动
5. 能够监控进程，并且当它们被异常停止的时候重启它们
6. 当发生硬件或软件的故障时，可减少downtime
7. 软件维护时，可以缩短计划的停机时间（downtime）

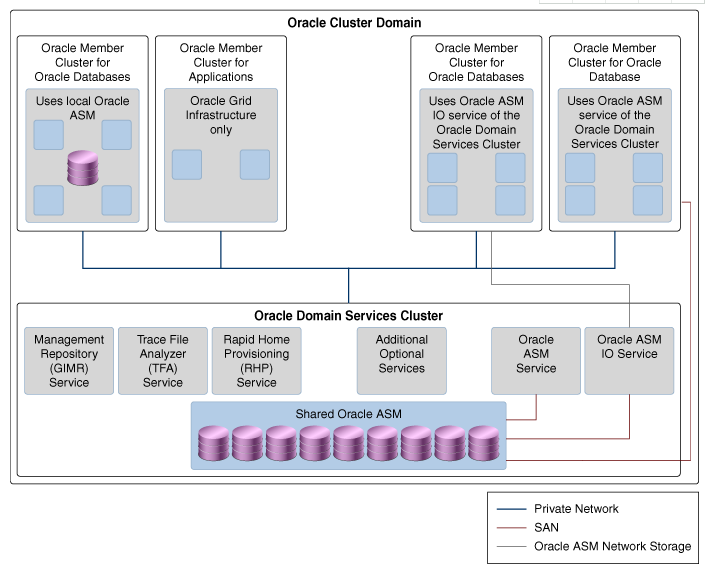
Oracle Clusterware自动管理的资源分俩类：cluster resource和local resource。

Oracle Clusterware有俩个必要的存储组件，都是二进制文件。一个是voting files，记录节点成员信息，一个就是OCR（Oracle Cluster Registry），记录集群配置信息。这俩个文件必须都放在共享存储上，所有节点必须都可以访问的到。

### Deployment Architectures

Oracle Clusterware提供的deployment architecture如下：

1. Domain services cluster:   
   A cluster that provides centralized services to other clusters within the Cluster Domain. Services can include a centralized Grid Infrastructure Management Repository (on which the management database for each of the clusters within the Cluster Domain resides), the trace file analyzer service, an optional Rapid Home Provisioning service, and, very likely, a consolidated Oracle ASM storage management service.
2. Database member cluster:   
   A cluster that is intended to support Oracle RAC or Oracle RAC One database instances, the management database for which is off-loaded to the domain services cluster, and that can be configured with local Oracle ASM storage management or make use of the consolidated Oracle ASM storage management service offered by the domain services cluster.
3. Application member cluster:   
   A cluster that is configured to support applications without the resources necessary to support Oracle RAC or Oracle RAC One database instances. This cluster type has no configured local shared storage but it is intended to provide a highly available, scalable platform for running application processes.



## System Requirements

### 硬件要求

首先，一个节点要配置俩个网络，所以至少得有俩张网卡（two network interface cards）：一张做公网（public network），一张做私网（private network）。Oracle集群的节点间innterconnect网络（私网）不支持使用crossover cables。

那么，如果要达到高可用的目的，则至少4张网卡，2张绑一起做公网，2张绑一起做私网。

然后，就是共享存储。ORACLE支持的共享存储包括：Network File Systems (NFSs)、iSCSI、Direct Attached Storage (DAS)、Storage Area Network (SAN) storage和Network Attached Storage (NAS)。

如果想要达到高可用的目的，则需要做冗余。

最后，就是每个服务起最好有自己的一块本地磁盘，放操作系统文件和数据库软件。这样做的好处就是：

1. 高可用，一个节点坏了并不会影响到另一个节点的工作
2. 升级打补丁的时候可以使用Rolling upgrade，从而减少downtime。

### 操作系统要求

查mos，必须根据Oracle集群软件和数据库软件的版本来进行选择。

### 软件要求

就俩种文件：voting files和OCR，它们被统称为Oracle Clusteware files。

这俩文件必须放ASM或共享存储上，如果ASM的底层存储没做RAID之类的冗余的话，那一定要把这俩文件多放几个地方。所以，一般我们专门用3块磁盘做一个NORMAL冗余的磁盘组，只放OCR和voting files。

下面我们来分别介绍这俩文件。

#### Voting Files

Oracle Clusterware使用voting files来明确哪些节点是集群的成员。

Voting files如果放ASM上，那么它的数量会跟据磁盘组冗余来自动确定。

ORACLE建议最少3份voting file（这就是为什么用3块磁盘做一个normal冗余的磁盘组放voting files和OCR），最多5份，虽然最多可支持存放15个voting file。

#### Oracle Cluster Registry

Oracle Clusterware使用OCR存放和管理Oracle Clusterware控制的组件的信息，这些组件包括：racle RAC databases、listeners、virtual IP addresses (VIPs)、services和any applications。

OCR将配置信息存储在一个数据结构中的一系列key-value对中。

Oracle对OCR存放的建议和限制：

1. OCR最多放5个地方
2. 每个OCR存放的地方都必须能被集群中的所有节点访问到
3. 在不止一份OCR的情况下，可以在线替换某一份失败的OCR。
4. 可以用来更新OCR的工具有：OEM、Oracle Clusterware Control Utility (CRSCTL)、Server Control Utility (SRVCTL)、OCR configuration utility (OCRCONFIG)、DBCA

### 网络配置要求

#### GNS和DNS

#### SCAN

Single Client Access Name (SCAN)的ip地址最少1个，最多3个。

#### Manual Addresses Configuration

可以不是用GNS和DNS。手工进行地址配置。说白了就是配置/etc/hosts文件。

需要如下内容：

1. 每个节点一个public address和一个hostname
2. 每个节点一个VIP，VIP必须与自己对应的public ip是同一网段，VIP在集群建好之前是不应当被使用或者被ping通的。
3. 最多3个scan IP

## Platform-Specific Software Components

### Technology Stack

Oracle Clusterware由俩个独立的Technology Stack组成：

1. upper technology stack：由Cluster Ready Services (CRS) daemon (CRSD)固定（anchor）。
2. lower technology stack：由Oracle High Availability Services daemon (OHASD)固定。

这俩种technology stack各自有很多进程来协助cluster的运作。

#### CRS Technology Stack

CRS Technology Stack利用如下进程来管理各种各样的服务：

1. Cluster Ready Services (CRS)：  
   CRS是管理集群中的高可用操作的主要程序。  
   CRSD进程是根据OCR中的配置信息来管理集群资源的。其管理操作包括对资源的：启动、停止、监控和故障转移。  
   当资源状态发生变化时，CRSD进程会产生events。  
   如果RAC也安好了，那么CRSD还会负责监控database instance，listener等等，在他们失败的时候负责把他们自动重新启动。
2. Cluster Synchronization Services (CSS)：  
   CSS通过控制哪些节点是集群的member和通知集群中节点的加入或删除来管理集群配置。  
   cssdagent进程可以监控集群并提供IO隔离。该服务之前是由Oracle Process Monitor Daemon (oprocd)提供的。  
   一个节点上的cssdagent进程的失败可能会导致Oracle Clusterware重启该节点。
3. Oracle ASM：  
   为了集群和RAC管理磁盘。
4. Cluster Time Synchronization Service (CTSS)：  
   为Oracle Clusterware管理集群的时间
5. Event Management (EVM)：  
   一个将Oracle Clusterware创建出来的event进行发布的后台进程
6. Grid Naming Service (GNS)：  
   集群的域名解析
7. Oracle Agent (oraagent)：  
   负责支持Oracle的一些特定的需求和复杂的资源。  
   This process runs server callout scripts when FAN（Fast Application Notification）events occur。11g里叫RACG。
8. Oracle Notification Service (ONS)  
   与FAN events进行通信的负责发布（publish）和订阅的服务（subscribe）。
9. Oracle Root Agent(orarootagent)：  
   一个专门的oraagent进程，帮助CRSD管理root拥有的资源，比如网络和VIP。

CSS、EVM、ONS负责跟其他节点的其他集群组件进行通信，也是数据库、应用同Oracle Clusterware高可用组件主要的通信链接。同时这些后台进程还监控管理着database的操作。

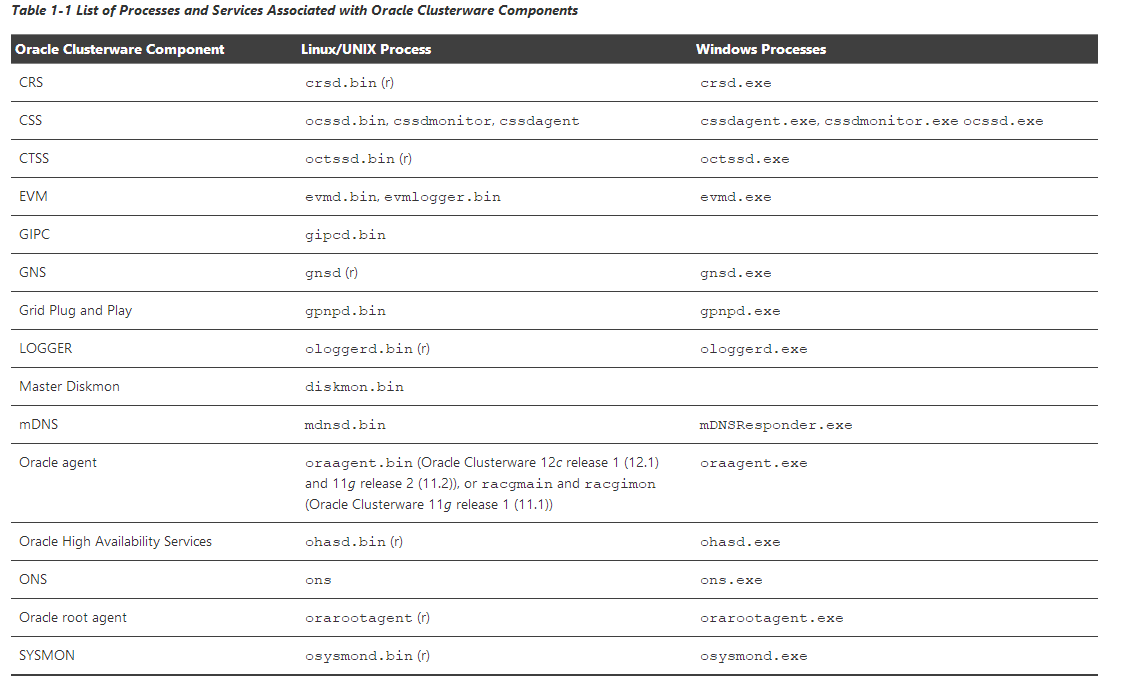
#### OHAS Technology Stack

OHAS（Oracle High Availability Services）利用如下进程来为Oracle Clusterware提供高可用：

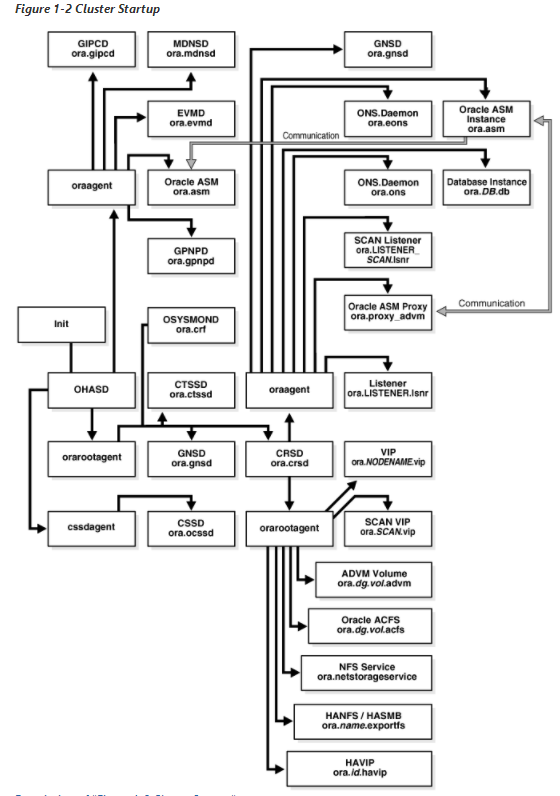
1. appagent：  
   包括之前版本的集群中的application资源类型的各种资源
2. Cluster Logger Service (ologgerd)：  
   接收集群中所有节点的信息，并把这些信息存在Oracle Grid Infrastructure Management Repository-based database中。这个服务只会在集群的俩个节点上跑。
3. Grid Interprocess Communication (GIPC)：  
   支持使用Redundant Interconnect Usage的一个守护进程
4. Grid Plug and Play (GPNPD)：  
   该进程提供对Grid Plug and Play profile的访问，同时协调各个节点对Grid Plug and Play profile的更新，使得所有节点访问到的都是最新的Grid Plug and Play profile。
5. Multicast Domain Name Service (mDNS)：  
   Grid Plug and Play用它来定位profile位置的进程，GNS用它来域名解析。mDNS是UNIX、LINUX和WINDOWS上才有的后台进程。
6. Oracle Agent (oraagent)：  
   管理GIPC、GPNPD、GIPC这些属于Oracle Clusterware的守护进程。
7. Oracle Root Agent (orarootagent)：  
   一个专门的oraagent进程，帮助CRSD管理root拥有的资源，比如Cluster Health Monitor (CHM)。
8. Scriptagent
9. System Monitor Service (osysmond)

### 集群启动流程

如下是集群相关进程的一个总结，（r）代表进程必须以root用户来运行：



如下是集群启动流程：



## 安装Oracle Clusterware概要

Oracle Clusterware 12c supports Oracle ASM 12c only, because Oracle ASM is in the Oracle Grid Infrastructure home, which also includes Oracle Clusterware

Oracle Clusterware 12c supports Oracle Database 12c and Oracle Database 11g release 2 (11.2.0.3 and later)

Oracle ASM 12c requires Oracle Clusterware 12c and supports Oracle Database 12c and Oracle Database 11g release 2 (11.2.0.3 and later)

Oracle Database 12c requires Oracle Clusterware 12c

Oracle RAC database上每个节点的ORACLE\_HOME必须一样。

Oracle Clusterware和Oracle Database的home所属的用户可以不一样，只要所属的primary group一样就行。

## 升级和打补丁概要

升级都得是out-of-place，打补丁（比如PSU）是in-place。Out-of-place意思就是换个GRID HOME，in-place就是就使用原GRID HOME。

## Grid Infrastructure Management Repository

Grid Infrastructure Management Repository（GIMR）就是存储集群的一些数据，例如实时性能数据、各种客户端采集和需求的元数据。

在之前的GI版本中，Grid Infrastructure Management Repository是GI home中的一个单独的数据库。需要使用第一个创建的磁盘组的磁盘空间，言外之意就是要跟OCR和voting files共享一个磁盘组。

现在，Oracle Member Clusters使用GIMR服务将其放于它的Domain Services Cluster中，也就是说它的GIMR现在就是Domain Services Cluster的GIMR CDB下的一个PDB。这样就不用跑一个local GIMR，并且与member cluster公用磁盘组了。（完全不懂说了个啥）

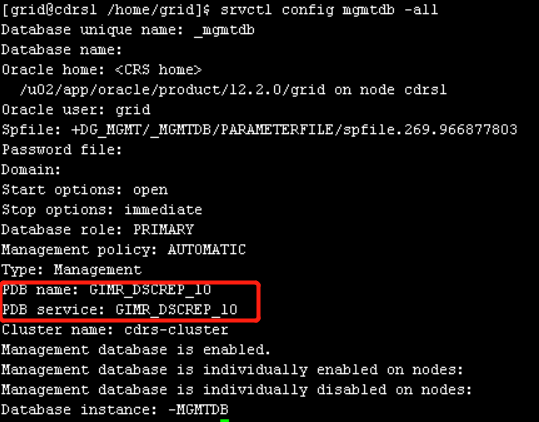
此外，我们可以在安装GI的时候，同时将local GIMR和Domain Services Cluster GIMR安装在MGMT磁盘组。注意domain services cluster所在的磁盘组绝不可以和OCR、voting files是同一个磁盘组。

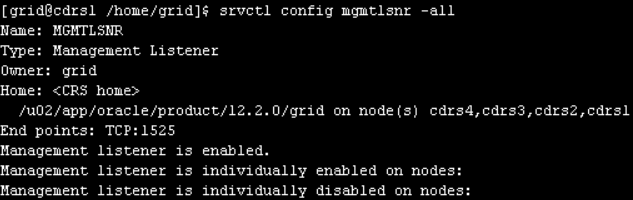
下面概述一下GIMR的功能：

1. 是一个存放由Cluster Health Monitor收集来的实时的操作系统的各种度量值的Oracle database
2. 可以在Oracle Clusterware 12c安装或者升级时配置GIMR
3. 必须运行在HUB nodes上，并且当磁盘失败或者节点失败时，必须能够failover到其他节点上
4. 与其他cluster clients (such as Cluster Health Advisor, cluster resource activity log, Rapid Home Provisioning Server, OLOGGERD, and OCLUMON)通过私网通信。
5. GIMR只会与external clients、Domain Services Clusters和member clusters通过公网进行通信
6. 强烈建议专门创建一个MGMT的磁盘组，供GIMR使用。否则，GIMR默认跟OCR、voting files用同一个磁盘组。
7. 可以通过srvctl中相关与mgmtdb、mgmtlsnr的命令进行管理

Srvctl config mgmtdb [-verbose] [-all]

Srvctl config mgmtlsnr [-all]





## Domain Services Clusters

## 管理集群环境概要

Oracle Clusterware提供给我们很多工具来管理集群环境：

1. Cluster Health Monitor (CHM)：  
   CHM不断的跟踪操作系统资源在node、process和device上的消耗。它不停的收集群数据，当发现超过某阈值、集群资源降级、节点被驱逐等等的问题时，会实时的给用户发出告警。  
   当发生失败后，我们还可以通过历史数据去追寻root-cause（根因）。
2. Cluster Verification Utility (CVU)：  
   CVU是一个命令行（command-line）工具。可以使用CVU来确认共享存储设备、网络配置、系统要求、Oracle Clusterware和操作系统的组和用户。  
   在集群和RAC安装前和安装后使用CVU进行预检查和确认是非常有用的。同时，对于一些集群的管理操作完成后，比如增删节点，也可以使用CVU来确认我们的配置是否正确。
3. Oracle Cluster Registry Configuration Tool (OCRCONFIG)：  
   OCRCONFIG是用来管理OCR的一个命令行工具。可以使用OCRCHECK和OCRDUMP工具来troubleshoot和OCR相关的配置问题。
4. Oracle Clusterware Control (CRSCTL)：  
   CRSCTL是管理Oracle Clusterware的命令行工具。  
   crsctl status resource可以监控和管理服务器（servers）和服务池（server pool）（不包括ora.\*开头的server pool），比如crsctl status server、crsctl status serverpool、crsctl modify serverpool、crsctl relocate server.  
   我们还可以使用crsctl start | stop | enable | disable | config has来管理整个集群的高可用服务，当然可以使用-n来指定某几个节点或者-all指定所有节点  
   使用crsctl start | stop | enable | disable | config crs可以管理特定节点上的Oracle Clusterware
5. Oracle Enterprise Manager：
6. Oracle Interface Configuration Tool (OIFCFG)：  
   OIFCFG是无论单实例数据库还是集群环境都可以使用的一个命令行工具。OFICFG可以为组件分配或者解除分配网络接口。  
   可以使用OIFCFG管理组件使用某个特定的网卡或者查看组件的配置信息。
7. Server Control (SRVCTL)：  
   可以管理Oracle资源，比如集群中的databases、services和listeners。还能并且也只能用SRVCTL管理ora.\*开头的server pool。

## Command Evaluation

使用CRSCTL可以评估一个crsctl操作的后果，而不是真实的执行。这个在基于策略管理的环境中很有用，可以帮助制定正确的策略。同样，对应ora开头的server pool，仍然无法使用crsctl进行评估，需要使用srvctl进行评估。

如下是官方给的例子：

$ crsctl eval delete server mjk-node2-3 -explain

Stage Group 1:

-----------------------------------------------------------------------------

Stage Required Action

-----------------------------------------------------------------------------

1 E Server 'mjk-node2-3' is removed from server pool 'sp1'.

E Server pool 'sp1' is below the MIN\_SIZE value of 2 with 1

servers.

E Looking at other server pools to see whether MIN\_SIZE value 2 of

server pool 'sp1' can be met.

E Scanning server pools with MIN\_SIZE or fewer servers in

ascending order of IMPORTANCE.

E Considering server pools (IMPORTANCE): sp2(2) for suitable

servers.

E Considering server pool 'sp2' because its MIN\_SIZE is 2 and it

has 0 servers above MIN\_SIZE.

E Relocating server 'mjk-node2-0' to server pool 'sp1'.

Y Server 'mjk-node2-3' will be removed from pools 'sp1'.

Y Server 'mjk-node2-0' will be moved from pools 'sp2' to

pools 'sp1'

===============================================================================

## Grid环境中Cloning and Extending Oracle Clusterware

如果想给新增节点，那么可以使用clone的方法。就是把现有已经装好集群软件的节点上的Oracle Clusterware software副本拷到新增的节点上，然后使用OUI等工具进行安装配置。添加或者删除节点会用到脚本addnode.sh和rootcrs.pl。

## Cluster Time Management

### 概述

Cluster Time Synchronization Service (CTSS)可以检测到集群节点之间的时间同步问题。

CTSS被作为Oracle Clusterware的一部分随集群一起安装。

当CTSS检测到系统上已经配置了例如NTP或者Chrony这样的时间同步服务时，不管这些服务是生效的还是无效的（broken），CTSS都会以observer模式运行。也就是说，如果检测到etc目录下有ntp.conf，那么不管NTP服务是否真的启动，CTSS都会运行在observer模式。

如果CTSS发现集群中的任何一个节点上都没有时间同步服务，那么CTSS会以active模式运行，接管集群的时间管理任务。

当CTSS运行在active模式时，一个非NTP的时间同步服务启动了，那么如果我们想让CTSS进入observer模式，可以直接在etc目录下创建一个ntp.conf文件。那么CTSS会记录一条信息在告警日志中，然后转为observer模式。

### step time synchronization

当新节点加入集群，并且CTSS是active模式，那么CTSS会以当前集群中的一个节点的时间为基准，比对新节点的时间。如果发现差异，并且该差异在可调整范围内（within a certain stepping limit），那么CTSS就会进行step time synchronization，通过把新节点的时间推前或者推后调整成基准时间。

### slew time synchronization

当集群中现有节点的时间与基准时间（第一个加入集群的节点的时间）不一致时，CTSS会进行slew time synchronization，就是加快或者放慢不一致节点上的系统时间，来最终使其与基准时间一致，这种方法跟step time synchronization不同，系统时间不会回退，保证了时间是单调递增的，通过调节时间增长速度，来使节点于基准时间保持一致。

如果Oracle Clusterware启动时，CTSS是active模式，发现节点与基准时间差异超过24小时，则CTSS会记录一条告警，然后集群启动失败。我们必须手动调整时间，能让集群顺利启动，然后由CTSS进行进一步的精准调整。

当使用slew time synchronization调整时间时，CTSS会定期向告警日志中写入将时间调整为基准时间的频次。

CTSS会在如下情况向Oracle Clusterware的告警日子和syslog中记录内容：

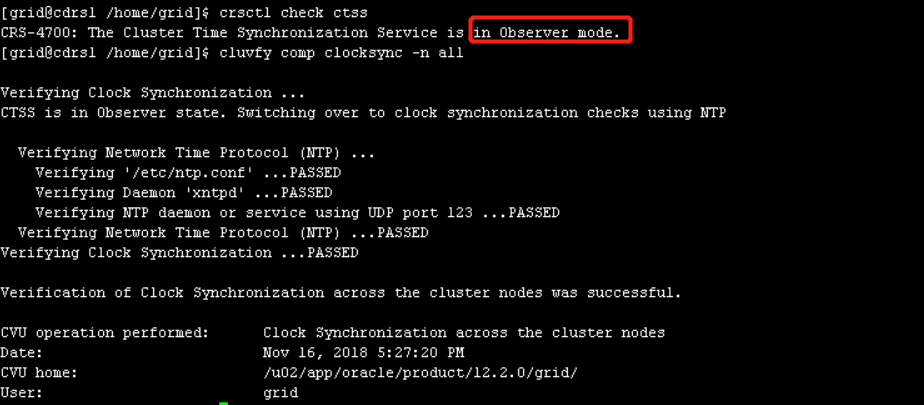
1. 发现时间改变
2. 发现节点时间与基准时间有明显差异
3. 由active模式转换为observer模式

### Active and Deactive CTSS

Active CTSS就是把所有节点上的第三方的时间同步服务停掉，并且把配置也要删掉，比如停了NTP，还要把etc/ntp.conf删掉。当CTSS检测到这些都没有的时候，会接管集群的时间管理。

Deactive CTSS就很简单：

1. 在所有节点配置第三方时间同步服务，CTSS会自动转为observer
2. 使用crsctl check ctss来确定CTSS已是observe模式
3. 启动所有节点上配置好的第三方时间同步服务
4. 使用cluvfy comp clocksync -n all来确定第三方时间同步服务已经正常工作



# 集群的配置和管理

## Role-Separated Management

角色分离管理是一种管理集群资源和工作负载的方法，其目的就是减少资源的争用或短缺。

角色分离管理使用operating system security、role definitions和Oracle Clusterware access permissions来根据用户的角色将资源以及工作负载分离管理。

注意，该特性并不会在安装过程中实施。

其实配置角色分离就是创建不通的操作系统用户和用户组，来分别管理集群资源（例如database），必要的时候，可以通过ACLs（access control lists）来添加对集群资源和server pools的访问限制。此外，ASM提供了进一步扩展角色分离架构的能力，就是创建可以控制对存储的访问的角色（比如asmadmin、asmdba）。

默认的，安装了Oracle Clusterware的用户和root用户是永久的cluster administrators。Primary group权限（默认是oinstall）让DBA在新建的server pool中建库。

### Managing Cluster Administrators

首先需要说明的是，只有cluster administrators才有创建server pool的权限。

那么在之前的版本中，默认将每一个注册了的操作系统用户都作为cluster administrators，所以当要限制这些用户的权限时，需要通过命令crsctl add | delete crs administrator来增删cluster administrators。

那么这个命令在12.2已经被废弃了，12.2中我们应该使用policy的ACL属性来控制创建server pool的权限。

如果一个用户或一个组想要创建server pool或者集群资源，那么它必须在ACL属性中的配置是可读、可写、可执行的权限。

### Configuring Role Separation

最基本的就是创建属于oinstall用户组的俩个用户（grid、oracle），然后给grid用户管理server pool的权限，其他用户对server pool只读的权限。

注意role separation的管理方式只适用于server pool和user-defined cluster resources and types，不能用于ora.\*的资源（Oracle RAC database resources）。

官方给的例子，对于一个叫psft的server pool，给personnel用户组下的personadmin用户对psft有rwx的权限，该用户组下的其他member对psft是r-x权限，其他用户组的对psft没有访问权限，那么要用root用户执行如下命令：

# crsctl setperm serverpool psft -u user:personadmin:rwx,group:personnel:rw-,other::---

再一个例子就是管理用户自定义的资源：

# crsctl setperm resource MyProgram -u user:Maynard:r-x,group:crsadmin:rw-,other:---:r--

## 使用Grid Setup Wizard配置GI

执行Oracle\_home/gridSetup.sh。

### Configuring a Single Node

主要分如下步骤：

1. grid用户下执行$ORACLE\_HOME/gridSetup.sh
2. 选择Configure Oracle Grid Infrastructure for a Cluster
3. 集群节点信息页面，只选择本地node的hostname和对应的VIP名
4. 继续进行剩余步骤，然后点finish
5. 按提示执行root.sh

### Configuring Multiple Nodes

跟上面不同的就是在节点信息配置页面，把要配置的其他节点都填上并选上。

### Upgrading Oracle Grid Infrastructure

执行了gridSetup.sh后，进入图形界面，选择Upgrade Oracle Grid Infrastructure，全部信息配置完后，点finish，然后按提示执行rootupgrade.sh。

### Running the Configuration Wizard in Silent Mode

Grid用户下：

$ORACLE\_HOME/gridSetup.sh -silent -responseFile file\_name

然后按提示执行相应的脚本（root.sh和Grid\_home/gridSetup -executeConfigTools）。

## Server Weight-Based Node Eviction

我们可以配置Oracle Clusterware failure recovery mechanism。当发生私网失败时，决定终结或驱逐（evict）哪个节点。

就当发生脑裂的时候，集群基于权重来选择让哪些实例或哪些服务存活。

对于数据库实例和资源可以通过在srvctl add/modify database或者srvctl add/modify service命令中指定-css\_critical yes。

对于非ora.\*的资源，也可以使用crsctl add/modify resource命令，在其中指定-attr "CSS\_CRITICAL=yes"

给一个server配重，还可以使用crsctl set server，指定-css\_critical yes。

有如下俩点额外注意的：

1. 配置了权重后，想要生效，需要重启集群或资源
2. 配置了权重后，管理模式由administrator managed变为policy managed，则配置失效，需要重新配置，当再次切回administrator managed后，配置重新生效。

## Oracle Database QoS Management

## Grid Naming Service（GNS）

## Node Failure Isolation

## 理解手动配置网络中的网络地址

### 理解网路地址配置要求

首先，至少得有俩张网卡，一张做公网，一张做私网。

我们可以配置一个公网网络使用IPv4协议或者IPv6协议或者俩个一起使用。但是，如果做网卡绑定的话，被绑定的网卡必须使用同一种协议。

我们可以配置一个或多个私网网卡，但是所有的私网不能混用IPv4和IPv6，协议必须一致。

注意：You can only use IPv6 for private networks in clusters using Oracle Clusterware 12c release 2 (12.2), or later.

虽然公网可以混用IPv4和IPv6，但是要么每个节点都混用，要么所有节点用同一个协议，说白了，每个节点的网络配置必须要一致，不能一部分值用IPv4，一部分值用IPv6。

#### IPv6的地址格式

IPv6地址格式如下，每一个x代表一个16进制字符：

xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx:xxxx

Oracle GI支持的IPv6地址如下：

1. RFC 4193中定义的Global and site-local IPv6 addresses（RFC 1884中定义的不支持）
2. IP地址中每一块区域中开头的0可以被省略
3. 压缩0位表示法：  
   我们可以把一个IPv6的地址2001:0db8:0000:0000:0000:8a2e:0370:7334写成2001:db8::8a2e:370:7334
4. 为了实现IPv4-IPv6互通，IPv4地址会嵌入IPv6地址中，此时地址常表示为：X:X:X:X:X:X:d.d.d.d，前96b采用冒分十六进制表示，而最后32b地址则使用IPv4的点分十进制表示，例如::192.168.0.1与::FFFF:192.168.0.1就是两个典型的例子，注意在前96b中，压缩0位的方法依旧适用

#### Name Resolution and the Network Resource Address Type

srvctl config network可以查看集群的网络配置

srvctl modify network -iptype可以修改IP类型

srvctl modify network -nettype可以配置地址的获取方式。IPv4的话可以是dhcp或static，IPv6的话可以是autoconfig或static。

如果是IPv4转IPv6的话，-nettype不能是mixed。

### 理解SCAN地址和Client Service Connections

我们知道公网是集群提供服务给客户段用的。

那么如果客户端连接服务器使用的是Single Client Access Name (SCAN)地址。那么当集群添加或删除节点时，虽然我们可能需要修改公网或私网的IP地址，但是对于客户端是无感知的，客户端并不需要修改他们连接配置。

Scan listener可以运行在集群中的任何一个节点上。在11.2以后，实例只会将scan listener注册位remote listener。

### SCAN listeners和使用valid node checking限制服务注册

## 手动改变IP地址

### 用SRVCTL改变VIP

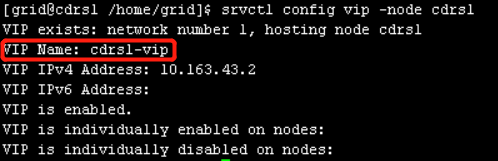
我们可以使用srvctl修改vip。修改过程如下：

首先，先把对外提供的服务先关掉。因为大概率应用配置连接该服务的tns使用的是vip。

srvctl stop service -db database\_name -service "service\_name\_list" -node node\_name

然后我们需要确定vip的信息：

srvctl config vip -node node\_name



再然后我们需要把要修改的节点上的vip停了：

srvctl stop vip -node node\_name

接着，我们使用ifconfig -a来确定vip是否真的停用了：

Ifconfig -a

再然后，修改/etc/hosts文件：

vi /etc/hosts

接着修改网络的子网（网段）、掩码、网卡名（要是都没变就不用修改，要是变了但没修改后面修改vip会报错），注意！这个一定要用root用户执行：

srvctl modify network -subnet subnet/netmask/interface

再接着就是该用srvctl修改VIP了：

srvctl modify nodeapps -node node\_name -address new\_vip\_address

这个new\_vip\_address的格式为：vip\_name|ip/netmask/[if1[|if2|...]]

例：srvctl modify nodeapps -node mynode -address 192.168.2.125/255.255.255.0/eth0

最后启动VIP还有最开始停掉的所有服务：

srvctl start vip -node node\_name

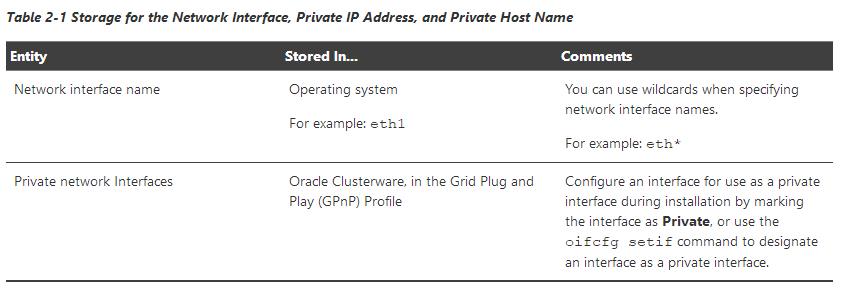
srvctl start service -db database\_name -service "service\_name\_list" -node node\_name

所有步骤完成后，可用如下命令，进行检查。该命令会列出所有可用的网卡，以及连接状态：

cluvfy comp nodecon -n all -verbose



### 修改私网配置（oifcfg）

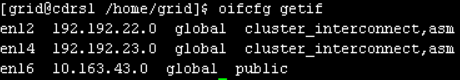


首先我们必须明确的是：配置私网时，所有节点必须使用同样的网卡名，并且在同样的子网（网段）中。

比如说node1的eth1和eth2作为私网网卡，那么node2的俩个做私网的网卡也必须叫eth1和eth2，并且所有节点的eth1必须在同一个网段，所有节点的eth2必须在同一个网段。

我们使用下面的命令查看，每个节点的输出结果必须都是一致的：

oficfg getif



### 用SRVCTL创建一个网络

使用root用户执行如下指令：

srvctl add nodeapps -node node\_name -address {vip |addr}/netmask[/if1[|if2|...]] [-pingtarget "ping\_target\_list"]

例：srvctl add nodeapps -node node1 -address node1-vip.mycluster.example.com/255.255.252.0/eth0

### 集群中的网络配置

就记住：

1. 每个节点的IP协议必须相同；
2. 进行网卡绑定的必须使用相同IP协议；
3. 做私网的网卡必须使用相同的IP协议

### IPv4和IPv6的转换

# Policy-Based Cluster

# Oracle Flex Clusters

# Rapid Home Provisioning, Scaling, Patching, and Upgrading

# 管理OCR和Voting Files

先介绍3个概念：

1. OCR：Oracle Cluster Registry管理集群和RAC数据库的配置信息
2. OLR：Oracle Local Registry，在每个节点上都有，管理自己节点的集群配置信息
3. Voting files：管理节点成员关系的信息。每一个Voting file都必须能被所有节点访问到。

## 管理OCR和OLR

管理OCR和OLR的工具有：OCRCONFIG、OCRDUMP和OCRCHECK。

这些管理工具的使用，必须是root用户。

我们知道OCR管理着所有Oracle resource的配置信息。

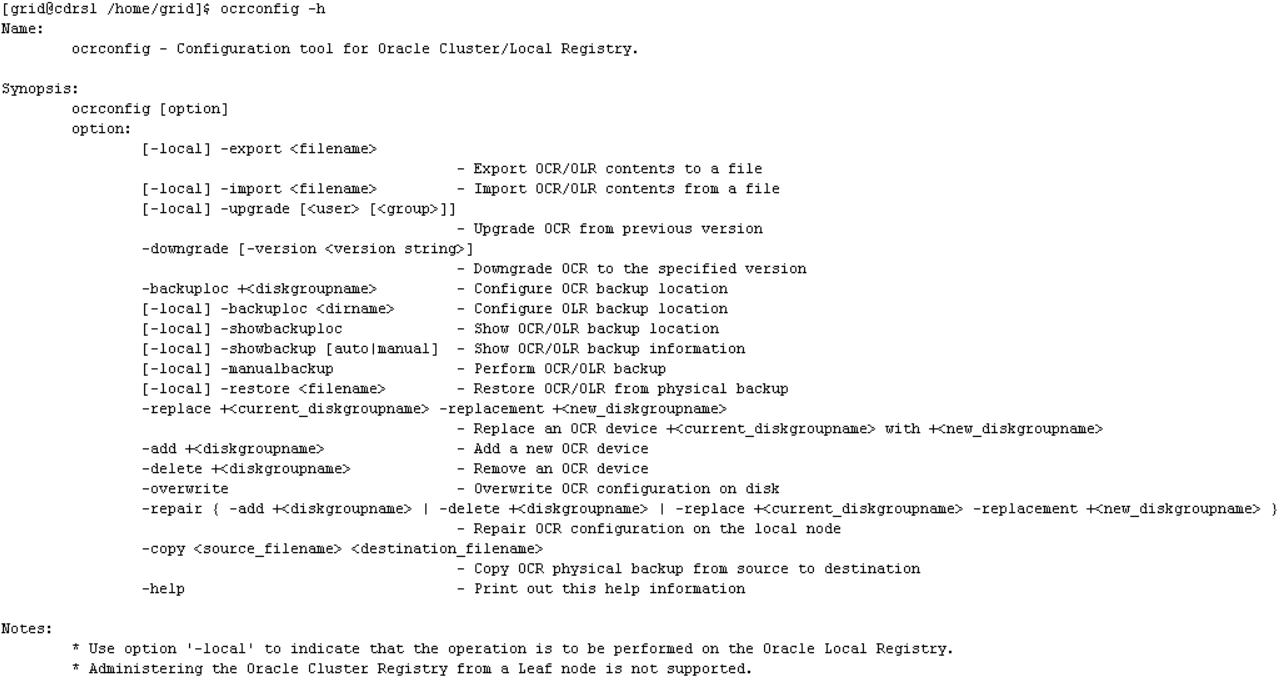
那么OLR的作用跟OCR差不多，只是，OLR只管理本节点的，所以OLR会放在本地（默认Grid\_home/cdata/host\_name.olr），而不像OCR要放在共享存储上。



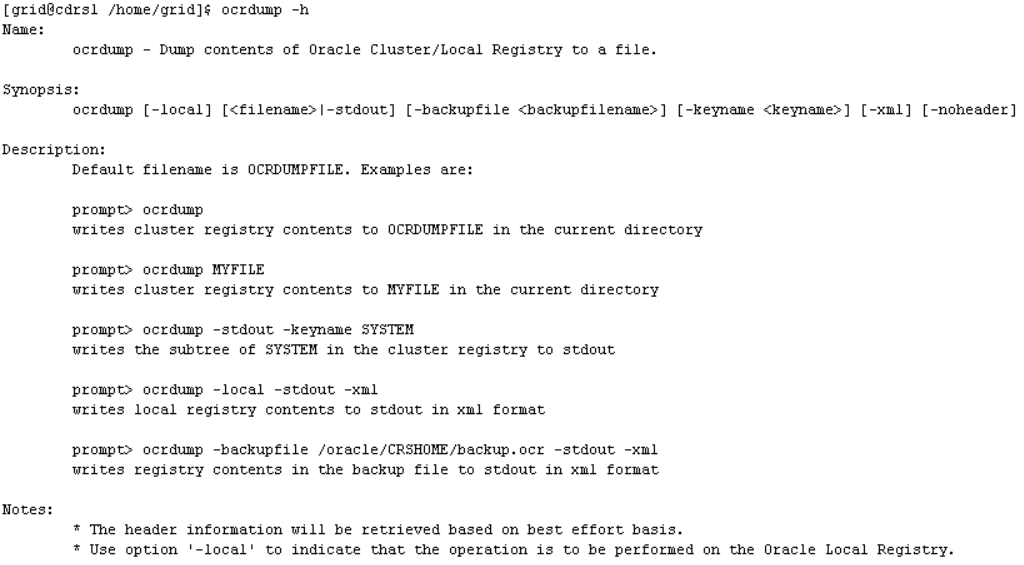
OLR会管理服务的依赖关系，OHAS就需要使用这些信息。

如下是这些工具的help：

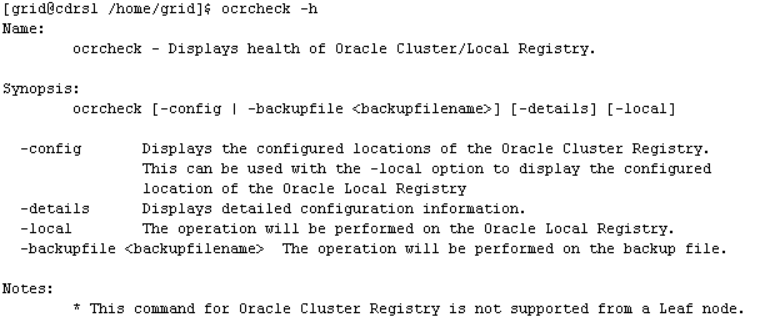
首先是OCRCONFIG：



然后是OCRDUMP：



最后是OCRCHECK：



### OCR

#### 迁移OCR到ASM

本节初衷是想说从老版本升级到新版本12.2的时候，需要将原本没有放在ASM上的OCR升级后放到ASM上。虽然这个需求不常有，但是我们可以学习，一些相关的命令。

首先，确定集群版本：

crsctl query crs activeversion



然后，把所有节点的ASM实例起起来，然后创建放OCR的磁盘组。

再然后使用root执行如下命令，将OCR放在指定磁盘组上：

ocrconfig -add +new\_disk\_group

注意，OCR最多只能放在5个地方，而且这5个地方的存储类型一定要相同。

最后把老地方的OCR删掉，也要用root执行：

ocrconfig -delete old\_storage\_location

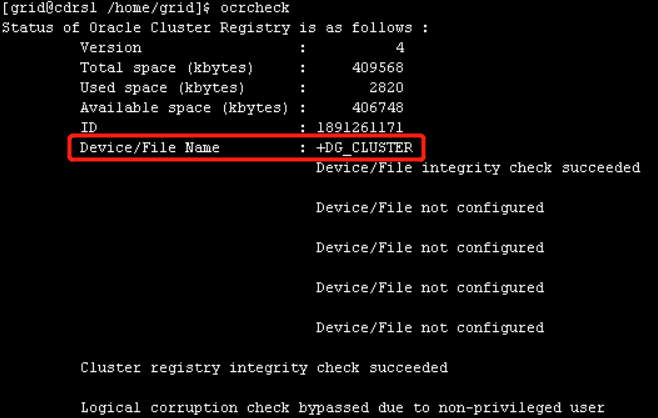
下面给出官文的一个示例：

# ocrconfig -add +OCR

# ocrconfig -delete /ocrdata/ocr\_1

最后，我们要注意，使用ocrchek检查OCR的配置：

ocrcheck



#### 从ASM移到其他类型的存储

依旧需要确认集群版本。

还要确认

1. 新的目录必须有root, oinstall, 640的权限
2. 新的存储地方必须有至少500M空间
3. 新的存储的地方必须对所有节点可见可访问

然后就是如下一些操作指令，跟往ASM上迁是一样的：

ocrconfig -add file\_location

ocrconfig -delete +asm\_disk\_group

再学一个新指令，改变OCR的备份路径：

ocrconfig –backuploc file\_location

#### Replace OCR的位置

其实前面的都是替换操作，那么如果直接执行ocrconfig -replace有一个前提条件：至少有俩个ocr location是online的。

如果只有一个ocr location是oline的，那么就只能使用ocrconfig -add和ocrconfig -delete的方法了。

指令格式如下：

ocrconfig -replace current\_OCR\_location -replacement new\_OCR\_location

#### CLSD-1009 and CLSD-1011

就是Oracle对OCR有一个保护机制。当突然一个节点上某个OCR location访问不到了，那么为了保证OCR的一致性，Oracle会禁止该节点上仍然可用的OCR location发生更新。并且会在告警日志中给出CLSD-1009和CLSD-1011的错误。

那么首先比对该节点上的OCR的配置（/etc/oracle/ocr.loc）跟别的节点的配置一样不。



然后如果不一样的话，用ocrconfig -repair修复

如果还解决不了的话，就用ocrconfig -overwrite来启动节点。

Ocrconfig -overwrite会使在错误发生时到现在这期间所有OCR发生的变化全部丢失。

#### 备份OCR

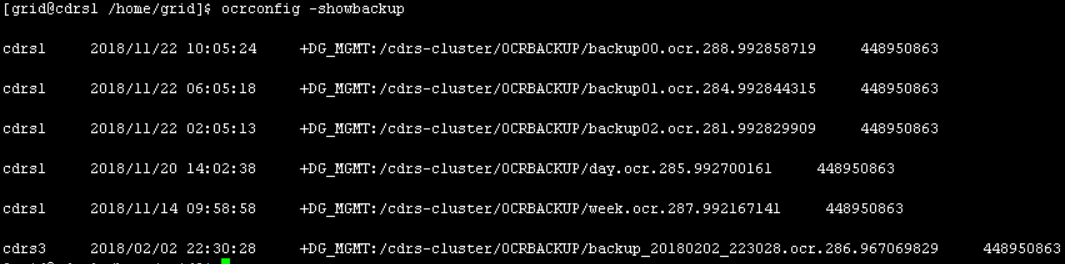
备份OCR分俩种：

1. Automatic backups：  
   这个是我们人为干涉不了的，我们无法设置备份间隔，也无法修改备份数量。  
   Oracle Clusterware每隔4小时创建一次OCR备份，并会保留最近创建的3份备份。  
   OCR备份的创建和保留的工作都是由CRSD进程完成。
2. Manual backups：  
   手工备份使用ocrconfig -manualbackup，执行操作的节点必须Oracle Clusterware stack是正常运行的。  
   一般在更改OCR之前，我们可以选择手动备份一下OCR。  
   注意，OLR只能通过手工备份。  
   但是，当在一个节点安装或升级集群，或者集群添加节点时，在执行完root.sh后，Oracle会自动备份OLR。

##### Listing Backup Files

使用命令ocrconfig -showbackup可以列出所有的ocr备份：

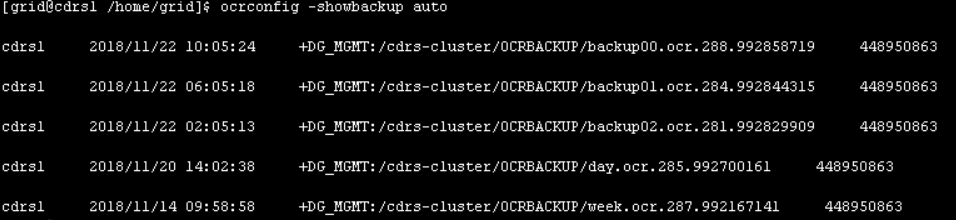
ocrconfig -showbackup [auto|manual]



我们可以看到：

1. 发起备份的节点
2. 备份的时间
3. 备份的路径
4. 备份的大小

默认是会把所有的备份全部列出来，我们也可以指定auto或者manual来只列出自动的或手动的备份：





我们可以使用如下命令来导出ocr备份中的内容：

ocrdump -backupfile backup\_file\_name



我们可以看到，会在当前路径下生成一个OCRDUMPFILE。里面的内容不是二进制的，是可读的。

##### Changing Backup Location

我们虽然无法控制OCR自动备份的时间间隔，但是我们指定备份的路径：

ocrconfig -backuploc file\_path

这里指定的路径，必须是所有节点都可以访问到的路径，如下是例子：

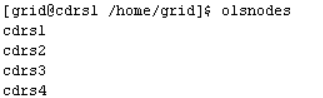
ocrconfig –backuploc +bkupdg

##### 恢复OCR

我们这里只记录Linux如何恢复OCR。

首先我们先看看有几个节点：

olsnodes



然后我们需要把crs先停了：

crsctl stop crs -f

直接加上-f，因为不加，会因为OCR已经损坏，而可能导致stop crs报错。

然后再把Oracle Clusterware stack启动到exclusive模式：

crsctl start crs -excl -nocrs

加-nocrs就是虽然启动Oracle Clusterware stack，但是不启动crs。

但我们还是要做一次检查，确认crs确实没有随着启动：

crsctl status resource ora.crsd -init

如果发现crs也跟着启动了，着执行如下指令停掉它：

crsctl stop resource ora.crsd -init

然后就该恢复OCR了。那么如果要恢复OCR，必须首先确定，OCR原来所在的磁盘组已经挂起。

也就是说，如果原本所在的磁盘组坏了，就需要先修复磁盘组，或者删了重建同名的磁盘组，如果要删的话要用force删：

SQL> drop diskgroup disk\_group\_name force including contents;

磁盘组准备好了，就进行OCR restore：

ocrconfig -restore backup\_file\_name

然后就是ocr检查，停掉把做restore的节点上的Oracle Clusterware stack：

ocrcheck

crsctl stop crs -f

剩余节点执行-repair -replace操作，比如我在2节点上做的restore操作，那么需要在1、3、4节点都执行-repair -replace操作：

ocrconfig -repair -replace

最后正常启动所有节点的crs，并进行检查：

crsctl start crs

cluvfy comp ocr -n all -verbose

#### Diagnosing Oracle Cluster Registry Problems

诊断就用ocrdump和ocrcheck诊断。

#### 用导出导入命令管理OCR

跟备份还原一个道理。

就是一样，在做一些关键性操作前（增删节点、升降级、创建数据库、修改资源），先导出OCR，就跟备份OCR一样，指令不同：

ocrconfig -export <filename>

就是注意，export出来的ocr，ocrconfig -restore不能使用，只有ocrconfig -import可以使用，反之亦然，backup出来的文件，import也不能使用。

官方建议使用backup备份OCR而不是使用export：

1. backup出来的是OCR的一致性snapshot，export出来的OCR块找如果想要一致，必须把所有节点停了再export。
2. 我们可以使用ocrdump来检查ocr备份的内容，但是export出来的文件我们无法查看其内容。
3. 用ocrconfig -showbackup可以列出OCR备份，但是想要知道export出来的文件放在哪儿，我们需要自己记录。

使用import恢复OCR方面，与之前不同的就是，不是用-restore，而是使用：

ocrconfig -import file\_name

其他都一样。

#### 升降级OCR配置

使用如下指令：

ocrconfig -upgrade

ocrconfig -downgrade

### OLR

每一个节点都有管理节点所属资源信息的local registry，它是在安装Oracle Clusterware时随OCR一同安装。

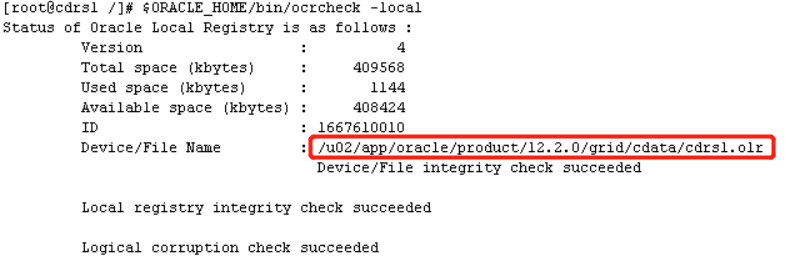
并且，不管Oracle Clusterware是否在运行，每个节点上都会有很多进程同时对OLR进行读写。

OLR的默认路径是：Grid\_home/cdata/host\_name.olr

我们说管理OLR也是使用ocrconfig、ocrdump、ocrcheck。与管理OCR不同的是，我们需要加上-local。

例如我们检查OLR的状态，root用户下执行：

Ocrcheck -local



如果我们想要查看OLR的内容，可以使用ocrdump -local：

ocrdump -local -stdout

使用-stdout，可以直接将结果输出出来，而不是dump到一个文件中。

想要export或者import OLR：

ocrconfig –local –export file\_name

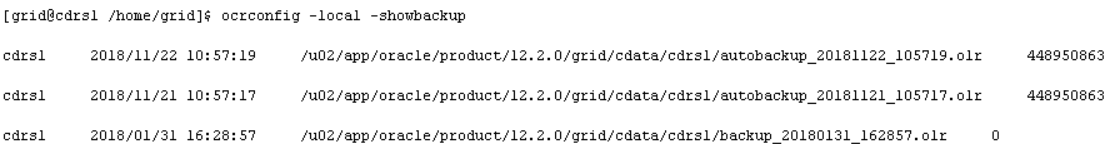
ocrconfig –local –import file\_name

想要手动备份OLR，查看备份，查看备份内容：

ocrconfig –local –manualbackup

ocrconfig -local -showbackup

ocrdump -local -backupfile olr\_backup\_file\_name



想要修改OLR默认备份路径并查看修改后的备份路径：

ocrconfig -local -backuploc new\_olr\_backup\_path

ocrconfig -local -showbackuploc



想要恢复OLR：

# crsctl stop crs

# ocrconfig -local -restore file\_name

# ocrcheck -local

# crsctl start crs

$ cluvfy comp olr

## 管理voting files

### Voting Files的存储

在同一个集群中，不可以把一部分voting files都放ASM上，另一部分不放在ASM上。

我们本节以voting files是放在ASM上的为前提。

Voting files有多少份取决于磁盘组的冗余级别，默认是一个failure group一份voting file。

冗余级别：

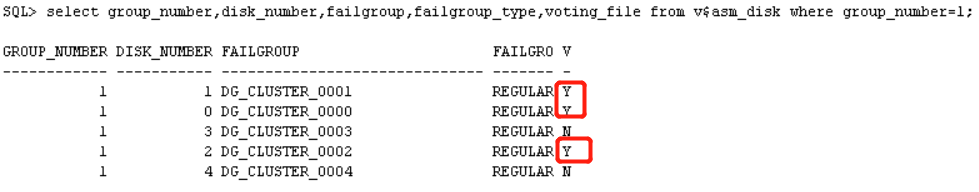
1. external就不说了
2. normal的话，普通磁盘组至少需要2个failgroup，也就是最少2块盘，而如果存放OCR+voting files的话，最少需要3个failgroup，会有3份voting files，1份OCR+1份OCR mirror
3. high的话，普通磁盘组至少需要3个failgroup，也就是最少3块盘，而如果存放OCR+voting files的话，最少需要5个failgroup，会有5份voting files，1份OCR+2份OCR mirror

可以使用crsctl replace votedisk把voting files从一个磁盘移到另一个磁盘组，通过改变磁盘组的冗余级别，来间接增加voting files的个数。

### 三个不能

不能直接干涉一个磁盘组中voting files的个数，因为它的多少取决于磁盘组本身的冗余级别。

我们来看个示例，即使是一个normal磁盘组中有5个failgroup，但是仍然只会有3份voting files：



不能对存放voting files的磁盘组使用crsctl add | delete votedisk，原因仍然是voting files的个数取决于磁盘组本身的冗余级别。也就是说，要是用ASM存放voting files，那么这俩命令就没用了。

不能在同一个集群下、在ASM磁盘组存在voting files的情况下，像非ASM的文件系统上增加voting files。

### 备份voting files

不用我们管，Oracle Clusterware会自动备份voting files的数据到OCR中，并且在新加入voting files时，将这些数据自动恢复到新的voting files上。

### 恢复voting files

如果voting files都坏了，那么我们就按如下步骤恢复。

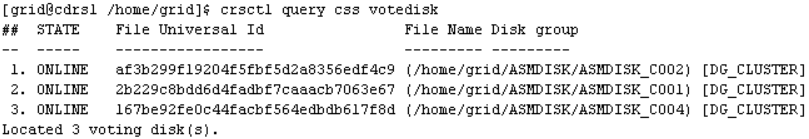
首先，OCR坏了要先按之前的步骤恢复OCR，如果没坏，那就没这步。

然后，在一个节点上，把crs启动到exclusive模式：

crsctl start crs -excl

再然后检查一下voting files，如果都损坏了，检查结果可能是空的，或者是status 3或者off：

crsctl query css votedisk



接着我们使用如下指令，将voting files移动到新的磁盘组：

crsctl replace votedisk +asm\_disk\_group

最后我们停crs，起crs：

crsctl stop crs

crsctl start crs

# 增删集群节点

# 克隆Oracle Clusterware

# Making Applications Highly Available Using Oracle Clusterware

本章将介绍如何使用Oracle Clusterware启动、停止、监控、relocate应用。

## Oracle Clusterware Resources and Agents

本节将讨论Oracle Clusterware使用怎样的framework来监控和管理资源，从而保证集群的高可用。

### Oracle Clusterware Resources

### Oracle Clusterware Resource Types

### Agents in Oracle Clusterware

### Action Scripts

### Building an Agent

### Registering a Resource in Oracle Clusterware

## Overview of Using Oracle Clusterware to Enable High Availability

# Troubleshooting Oracle Clusterware