# 1 HA集群介绍

Messaging Layer

M

D

B

A

C

高可用集群就是:当一个节点发生故障时，另一个节点能够发现并顶替该节点对外服务最少有两个节点。

每个节点都有一定的票数，当节点无法通信时，会根据分组票数相加计算，若某组票数过半则认为其他组发生故障，并将其隔离。另外，故障组经过计算后，发现自己票数不足一半，一般就会采取三种行动

freeze 不接受新的请求

stop 停止服务

ignore 继续运行服务

高可用集群标准：

HTA=MTTF/（MTTF+MTTR）

99% 一年宕机时间不超过4天

99.9% 10小时

99.99% 一小时

99.999% 6分钟

## 1 资源

资源就是启动一个服务需要的子项目，例如当启动httpd时，需要固定的ip，需要服务（包含软件 服务脚本 配置文件 甚至网页文件等），还需要文件系统，这些我们都可以称为资源。

**资源类型：**

**clone或者master/slave必须先是主资源或者组资源后。然后才能被成为……**

**clone是主资源复制而来。**

primitive 主资源，仅能运行在一个节点上

clone 能同时运行在多个节点上，如爆头设备的进程，分布式锁管理器（集群文 件系统自带），集群文件系统的锁对所有节点可见。

group 把多个主资源归类，同进同退

master/slave 特殊的clone克隆资源，仅仅能运行在两个节点上。drbd上常见

## 2 Message Layer：

传递当前节点的心跳信息，并告知对方。

可以提供该组件的软件有：

heartbeat常见/corosync/keepalive/cman/heartbeat

## 3 CRM：Cluster Resource Manager

集群资源管理器，需要借助Message Layer来实现工作，因此工作在Message Layer的上层，主要工作就是根据Message Layer传递的健康信息决定服务的启动 停止 和转移。每个节点上都要运行一个CRM，且每个CRM都维护着一个CIB（cluster Internet base）集群信息库，只有主节点上的CIB是可以修改的，其他节点上的CIB都 是在主节点上复制而来的。

常见的CRM为

haresources

crm

pacemaker 最常用。

rgmanager

**CRM之间通过UDP协议进行传输**，效率高，且因为位于同一局域网，因此安全性 也较高。

该程序包含了以下部分：

**DC**：Designated Coordinator 事务协调员，当多个节点彼此收不到对方的心跳信息，各个节点都会认为对方发生故障，于是就会产生分裂状况（分组），就会产生资源争夺的状况。因此事务协调员诞生。他根据每个组的法定票数来决定哪些节点启动服务，哪些节点关闭服务。DC一般位于主节点上。当某组法定票数过半，但是DC并不在该组中，该组就会推选DC。

PE：Police Engine 策略引擎，判断是否转移等，只是决策，并不执行

TE：Transaction Engine 事务引擎，执行PE做出的策略

LRM：Local Resource Manager 负责获取某个资源的状态，并且管理本地资源。

## 4 RA资源代理

**LRM是管理本地资源的，但是不能启动资源，当需要启动资源时会调用RA来启动，**RA是一个脚本文件，在一个节点上可能会有多个RA。通常在Linux下，启动系统服务不也都是一些脚本文件么。常见的RA有如下风格：

LSB Linux Standard Base：这是我们常见的如/etc/init.d/下的标准Linux脚本风格

OCF Open Cluster Fram：比lsb更强大的一种脚本，支持多种参数

## 5 资源粘性：

资源是否更倾向运行于**当前**节点，正值表示资源倾向于该节点，负值表示资源倾向于逃离该节点，还有正无穷和负无穷。最终资源在哪个节点上取决于资源粘性和资源的位置约束。

例如：当node1资源粘性为100，而资源对node2的位置约束为正无穷，则资源倾向于node2。因为资源对当前节点（也就是node1）的粘性为100，而资源二则是位置约束为正无穷。正无穷〉100.

## 6 资源约束 constraints

位置约束：location，根据score判断资源优先节点位置

正值：该资源倾向于此节点；inf，正无穷，必须在此节点

负值：该资源倾向逃离此节点，-inf，负无穷，必须不在此节点

顺序约束：order，定义资源启动或者关闭时的次序

排列约束：colocation(场地)根据score判断资源是否可以同时运行于同一节点

正值：可以在一起；inf，正无穷，必须在一起

负值：不可以在一起，-inf，负无穷，必须不一起

## 7 资源隔离

如果主动节点仍处在工作状态，但是被动节点不能接收到主动节点的心跳信息，就会立刻抢占资源，若这时主动节点正在向磁盘写入信息就会导致文件系统错乱或者崩溃，因此抢占资源的时候可以使用资源隔离的机制防止事故发生。

节点级别：stonithd shoot the other node in the head

被动节点直接操纵电源，控制故障节点的电源开关让其断电。

资源级别：

例子：FC SAN switch可以实现在存储级别上拒绝某节点的访问

## 8 共享存储

常用的httpd和mysql等服务需要将数据共享。常见的共享存储设备有如下三种：**下一节详细讲述了该信息**

DAS：直接附加存储

NAS：网络附加存储

SAN：存储区域网络