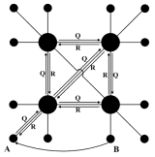
DATAGURU 炼数成金

Leader-us系列



大型分布式系统案例实战 第11周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

打造高可用的分布式系统（下）

- OpenAIS介绍
- Corosync入门
- Pacemaker入门
- Corosync +Pacemaker案例

OpenAIS介绍

资源、服务与主机(又称节点)的关系：

- 资源包括vip，httpd，filesystem等；
- 可整合多个资源形成一个服务；
- 服务必运行在某个主机上，主机上也可不运行服务(此为空闲主机)；
- 服务里的所有资源应该同时运行在同一个节点上

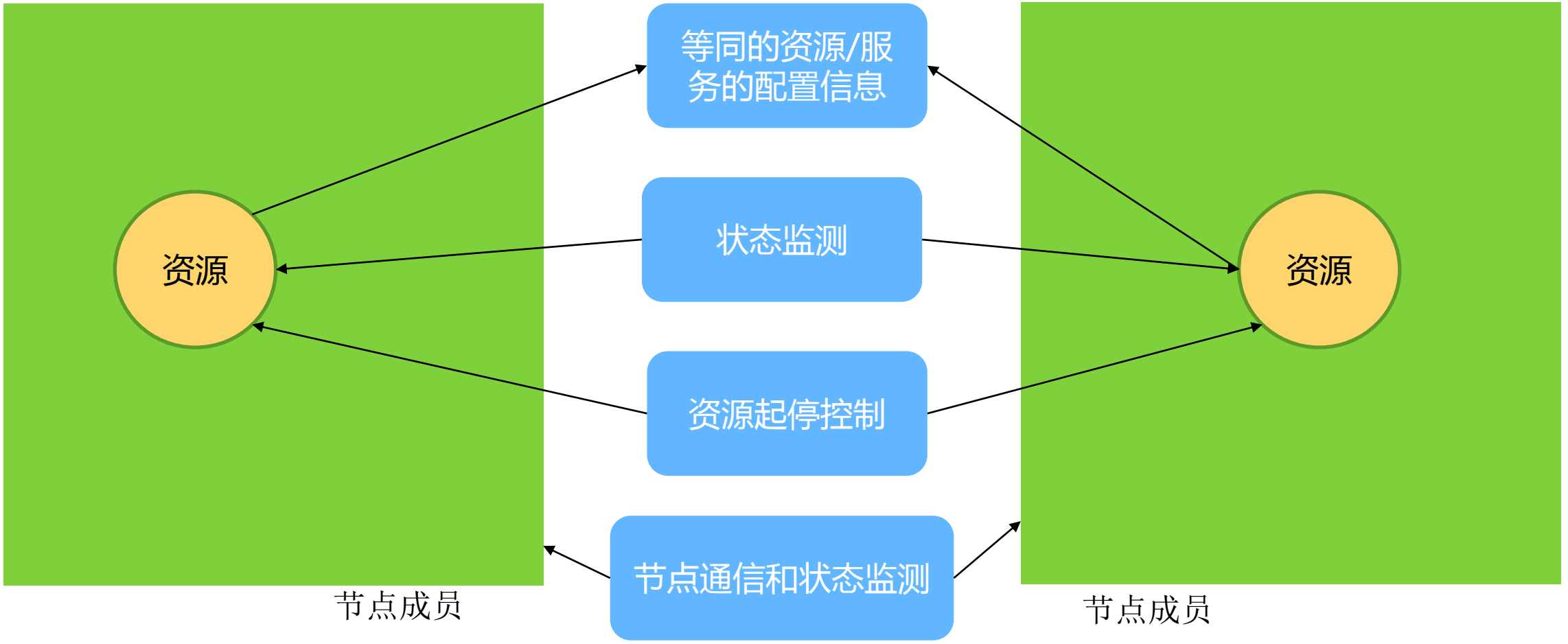
资源类型

- primitive(或native)：原生资源，只能运行于一个节点
- group：组资源
- clone：克隆资源，只能将原生资源定义为克隆属性；一般用于定义stonith设备的参数(可定义"总克隆数"和"每个节点最多可行的克隆数"等参数)
- master/slave：主从资源，也是克隆类型的，只能克隆2份；主的能读能写，从的不能读也不能写(如drbd的实现)

资源故障转移

资源黏性：资源运行在当前节点上是否远离的倾向性，数值越大表示越倾向留在当前节点
位置约束：资源更倾向于运行在哪个节点上，使用一个数值来表示约束的大小，数值越大越倾
排列约束：资源运行在同一节点的倾向性
顺序次序：定义资源的启动次序及关闭次序

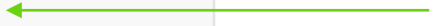
OpenAIS



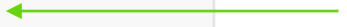
OpenAIS

OpenAIS提供一种集群模式，这个模式包括集群框架，集群成员管理，通信方式，集群监测等，能够为集群软件或工具提供满足 AIS标准的集群接口，但是它没有集群资源管理功能，不能独立形成一个集群

Project	branch	version	infrastructure
OpenAIS	whitetank	0.80.6	YES
OpenAIS	wilson	1.x	NO
Corosync	flatiron	1.x	YES
Corosync	needle	2.x	YES



接口API部分



基础设施和核心服务部分

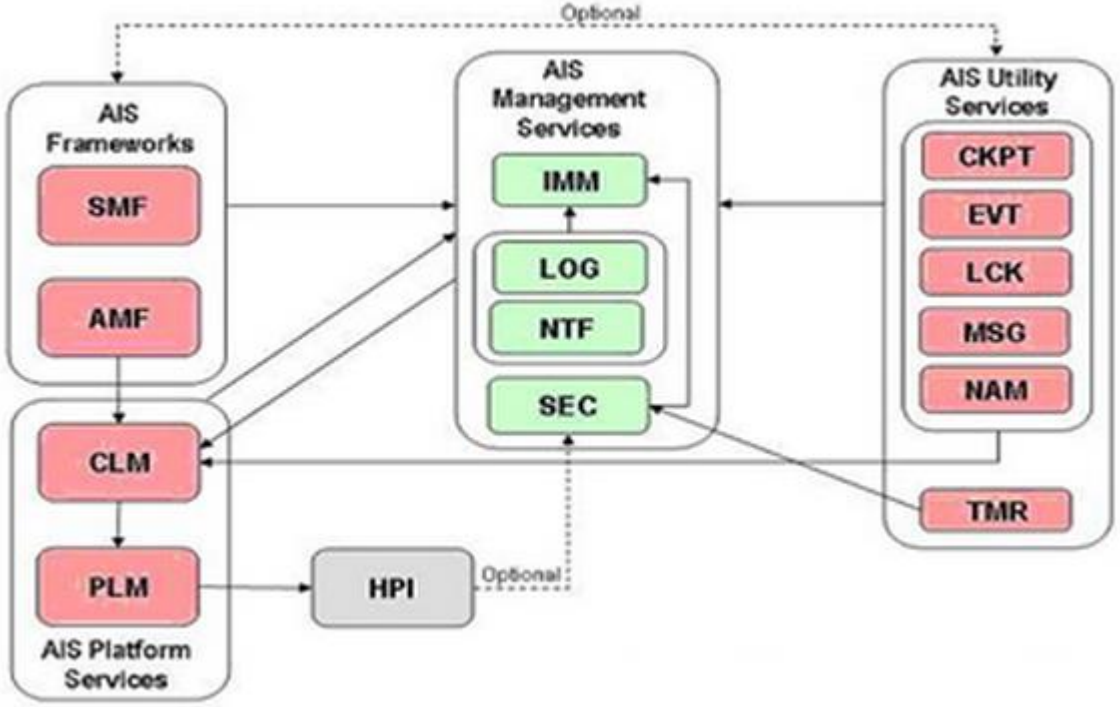
OpenAIS介绍

AIS Benefits:

- Reduced time to market and development costs
- Enhanced portability and integration capabilities
- Limits technology risk through choice of compatible COTS components
- Improved scalability for fault monitoring and management
- Increased resources focused on innovation of solutions

OpenAIS

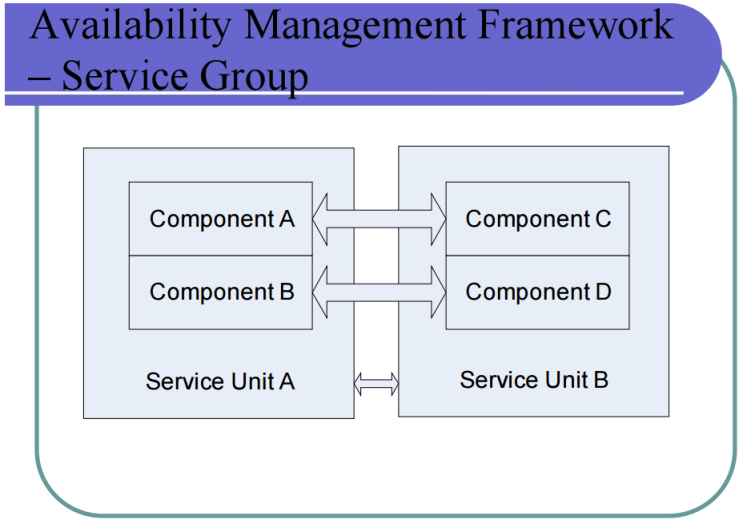
SMF（Service Mangement Framework)
AMF(Availability Management Framework)



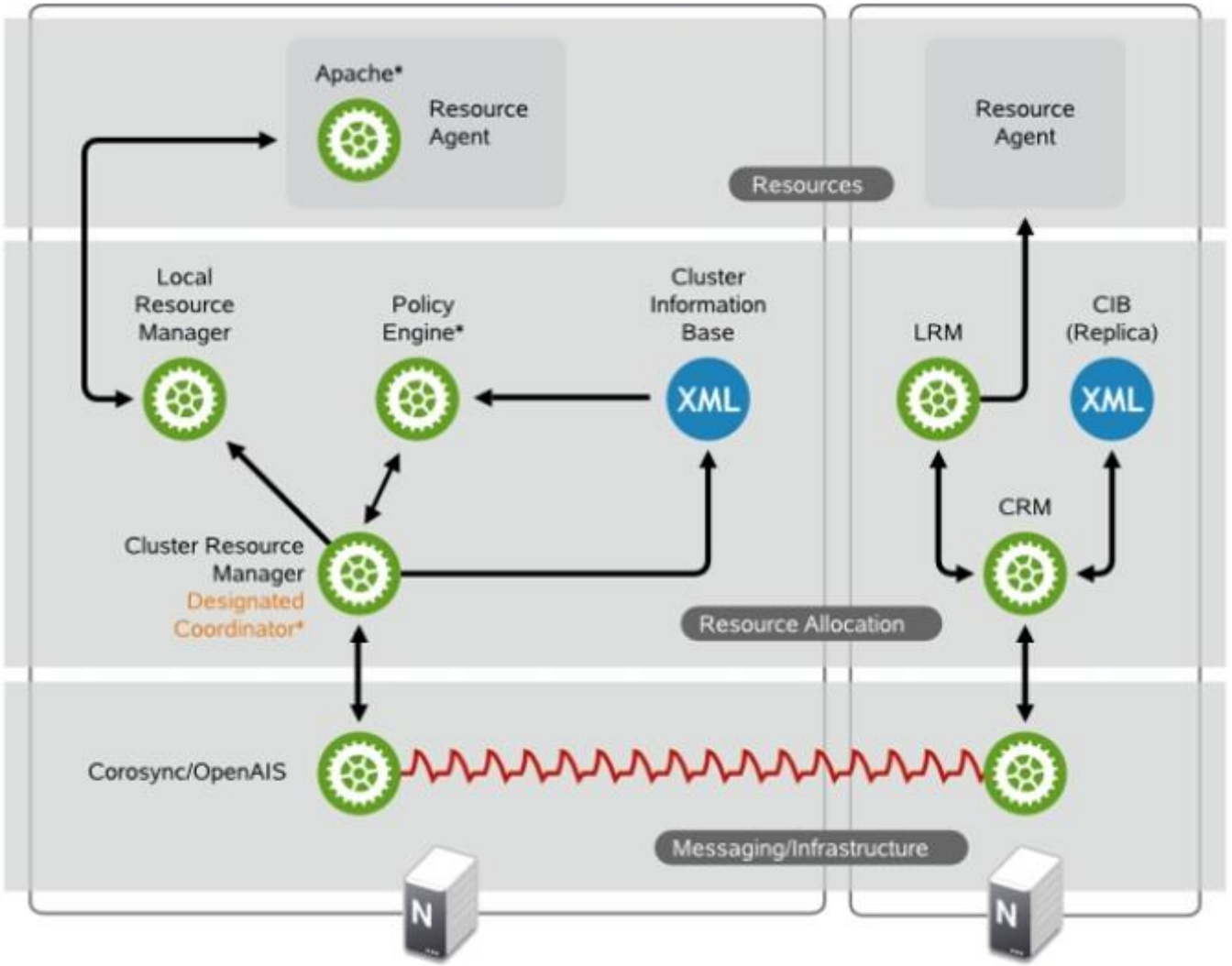
CLM(Cluster Membership Service)
PLM（The Platform Management Service）

IMM(Information Model Management Service)

CKPT(Checkpoint Service)
EVT(Event Service)
LCK(Lock Service)
MSG(Message Service)
TMR(Timer Service)



Corosync入门



资源代理层：RA是已写入的用来启动、停止和监视某种服务（资源）的程序

资源分配层，每一个动作的执行都要经过CRM。
在每一个节点上，CRM都会维护一个CIB。
某一个CRM被推选为DC。
LRM是CRM的代理，代表 CRM 调用本地RA.它可以执行start/stop/monitor操作并把结果反馈给CRM。

成员关系层，生产一个完整的成员关系

通过这一层发送“我还活着”的信号

Corosync入门

CRM中的几个基本概念

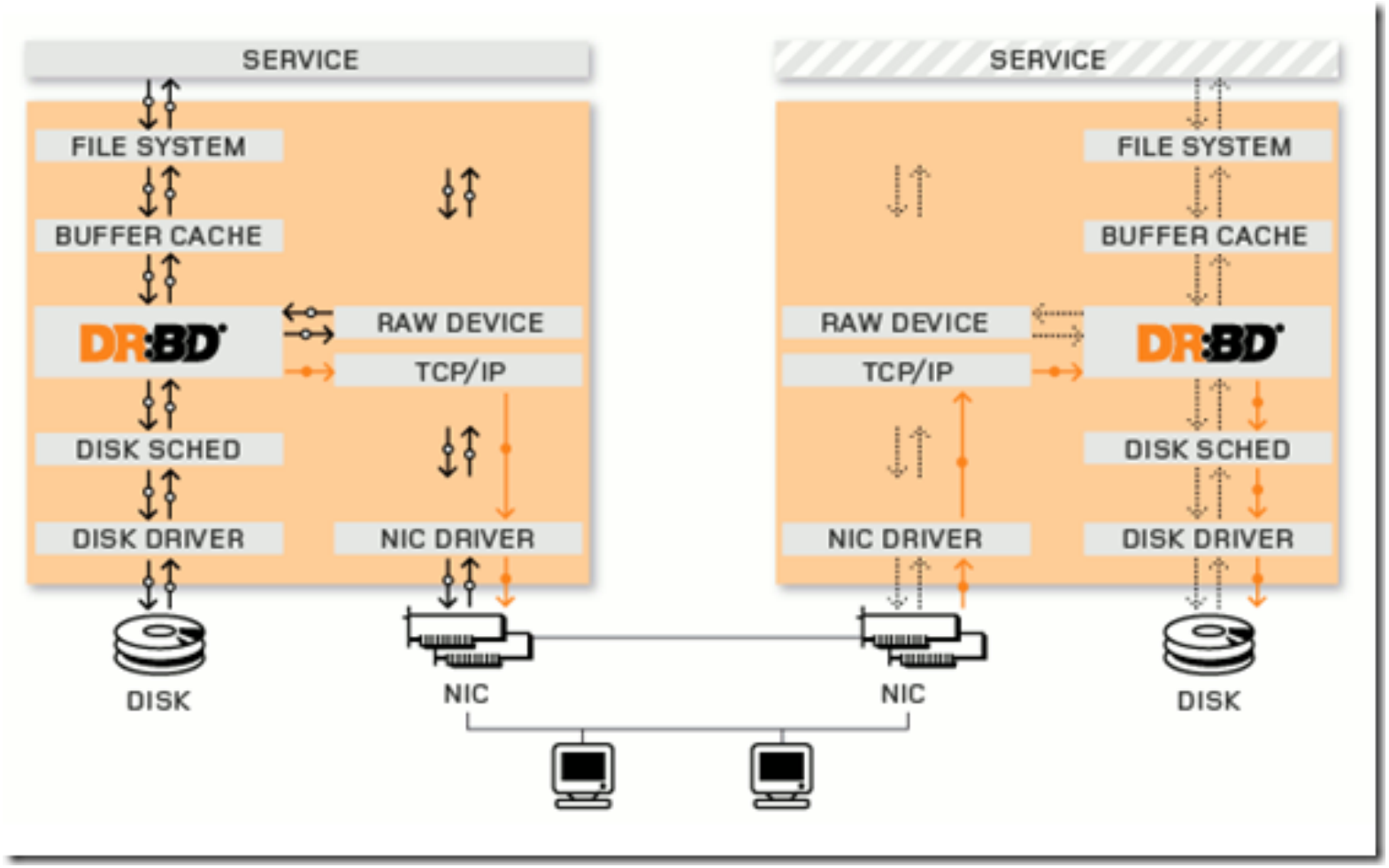
资源粘性：

资源粘性表示资源是否倾向于留在当前节点，如果为正整数，表示倾向，负数则会离开，-inf表示正无穷，inf表示正无穷。

资源类型：

- **primitive（native）：**基本资源，原始资源
- **group：**资源组
- **clone：**克隆资源（可同时运行在多个节点上），要先定义为primitive后才能进行clone。主要包含STONITH和集群文件系统（cluster filesystem）
- **master/slave：**主从资源，

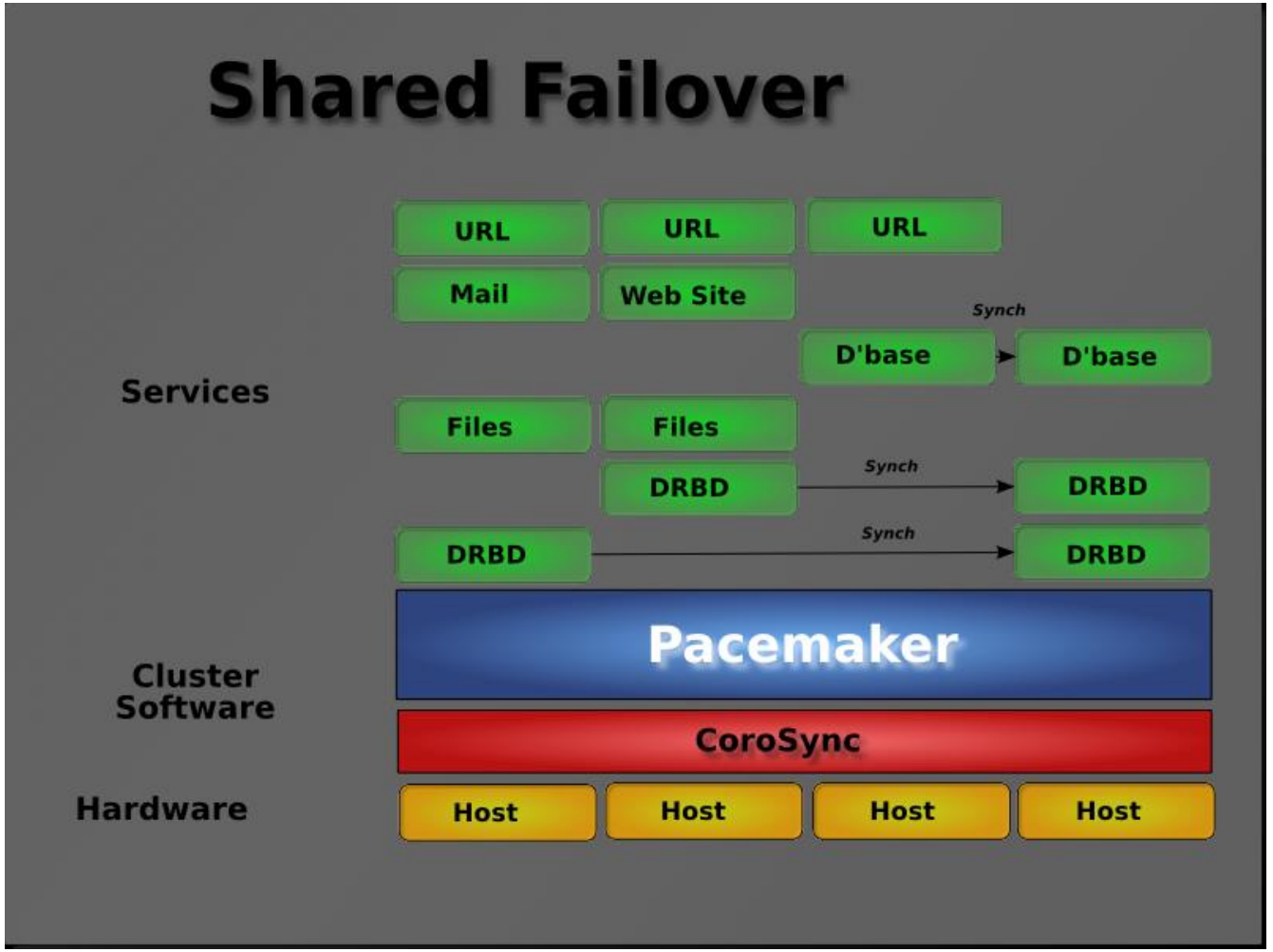
Corosync入门

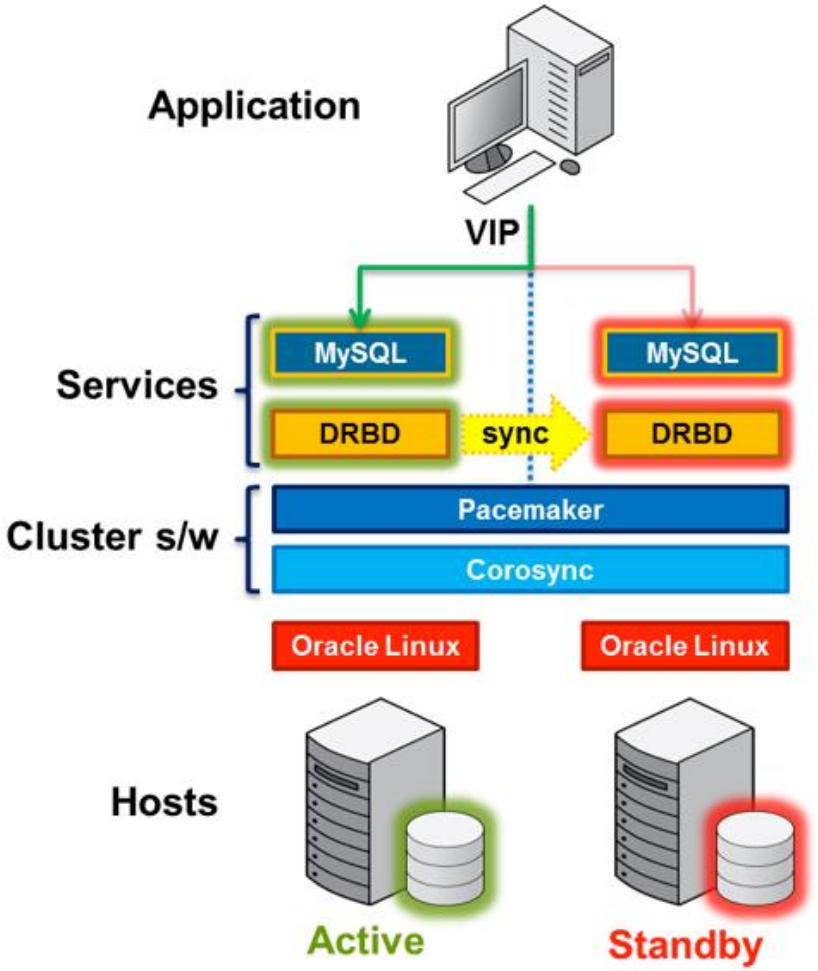


DRBD：Distributed Replicated Block Device 分布式复制块设备

DRBD有主双架构和双主架构的，当处于主从架构时，这个设备一定只有一个节点是可以读写的，另外的节点是不可读的，连挂载都不可能，只有一个节点是主的，其它节点都是从的。当做为主主架构时，需要达到几个条件，1.在高可用集群中启用DRBD； 2. 启用分布式文件锁功能，即需要把磁盘格式化为集群文件系统（如GFS2，OCFS2等）；3.把DRBD做成资源。

Corosync入门


Shared Failover
URL
URL
URL
Mail
Web Site
Synch
D'base
D'base
Services
Files
Files
Synch
DRBD
DRBD
Synch
DRBD
DRBD
Pacemaker
Cluster Software
CoroSync
Hardware
Host
Host
Host
Host


Application
VIP
Services
MySQL
MySQL
DRBD
sync
DRBD
Cluster s/w
Pacemaker
Corosync
Oracle Linux
Oracle Linux
Hosts
Active
Standby

Pacemaker入门

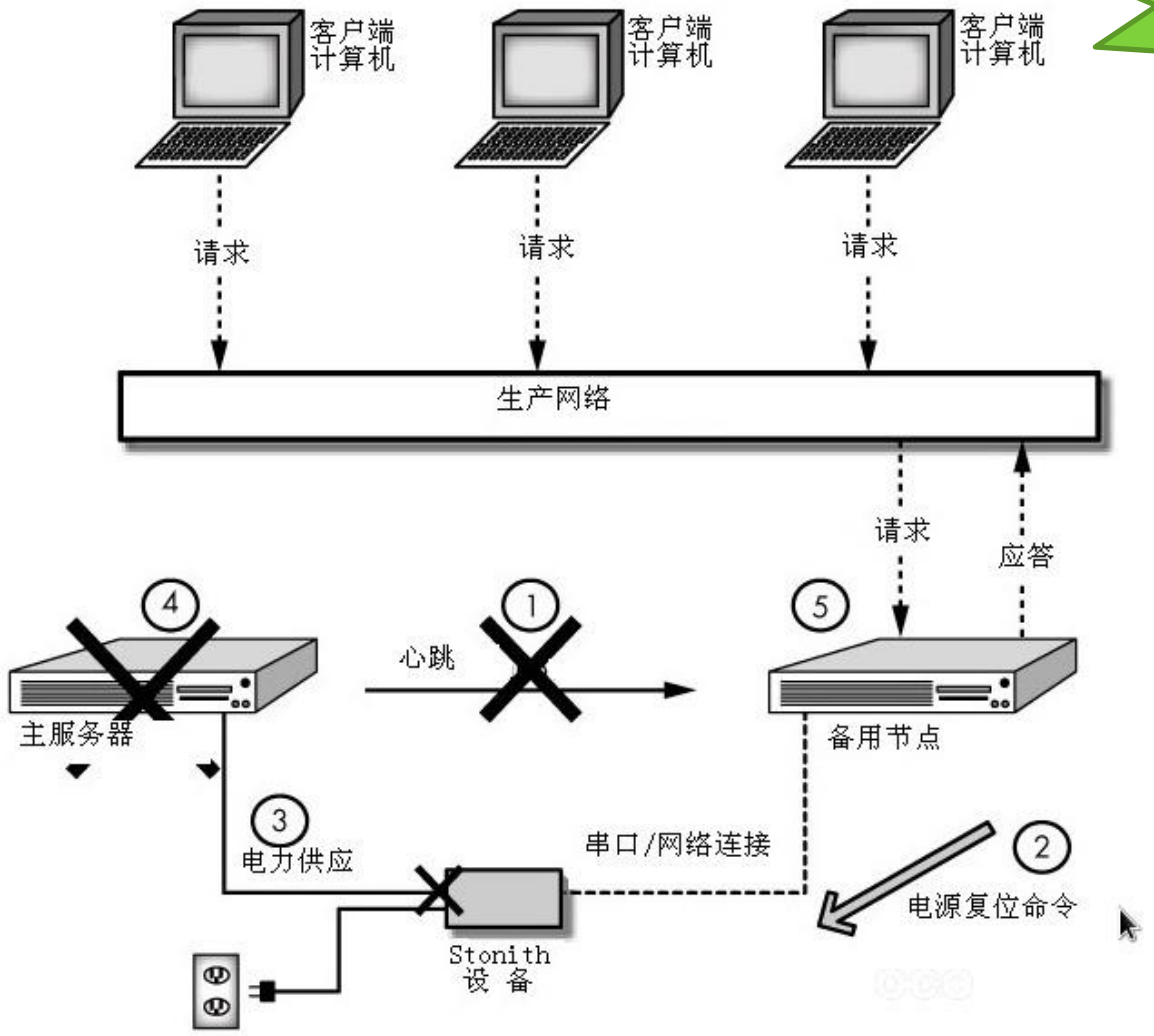
Pacemaker包含以下的关键特性:

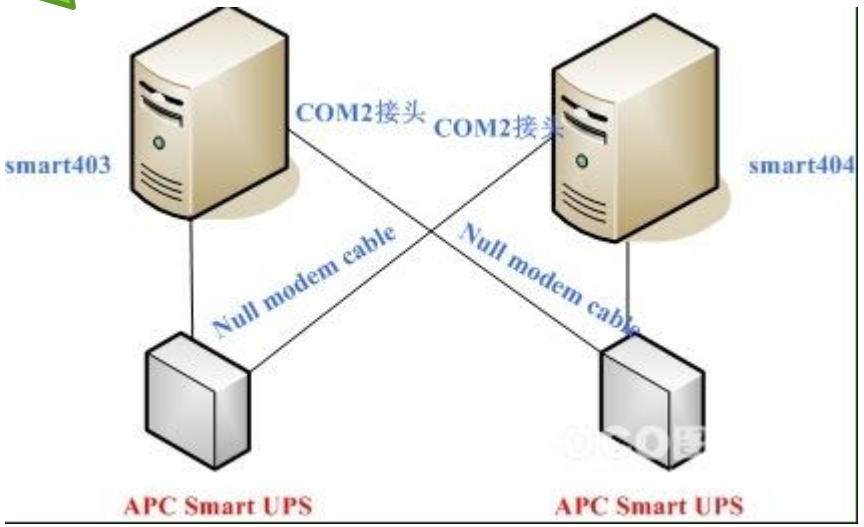
监测并恢复节点和服务级别的故障

- 存储无关，并不需要共享存储
- 资源无关，任何能用脚本控制的资源都可以作为服务
- 支持使用**STONITH**来保证数据一致性。
- 支持大型或者小型的集群
- 支持任何的 冗余配置
- 自动同步各个节点的配置文件
- 可以设定集群范围内的**ordering, colocation and anti-colocation**

Pacemaker入门

脑裂问题

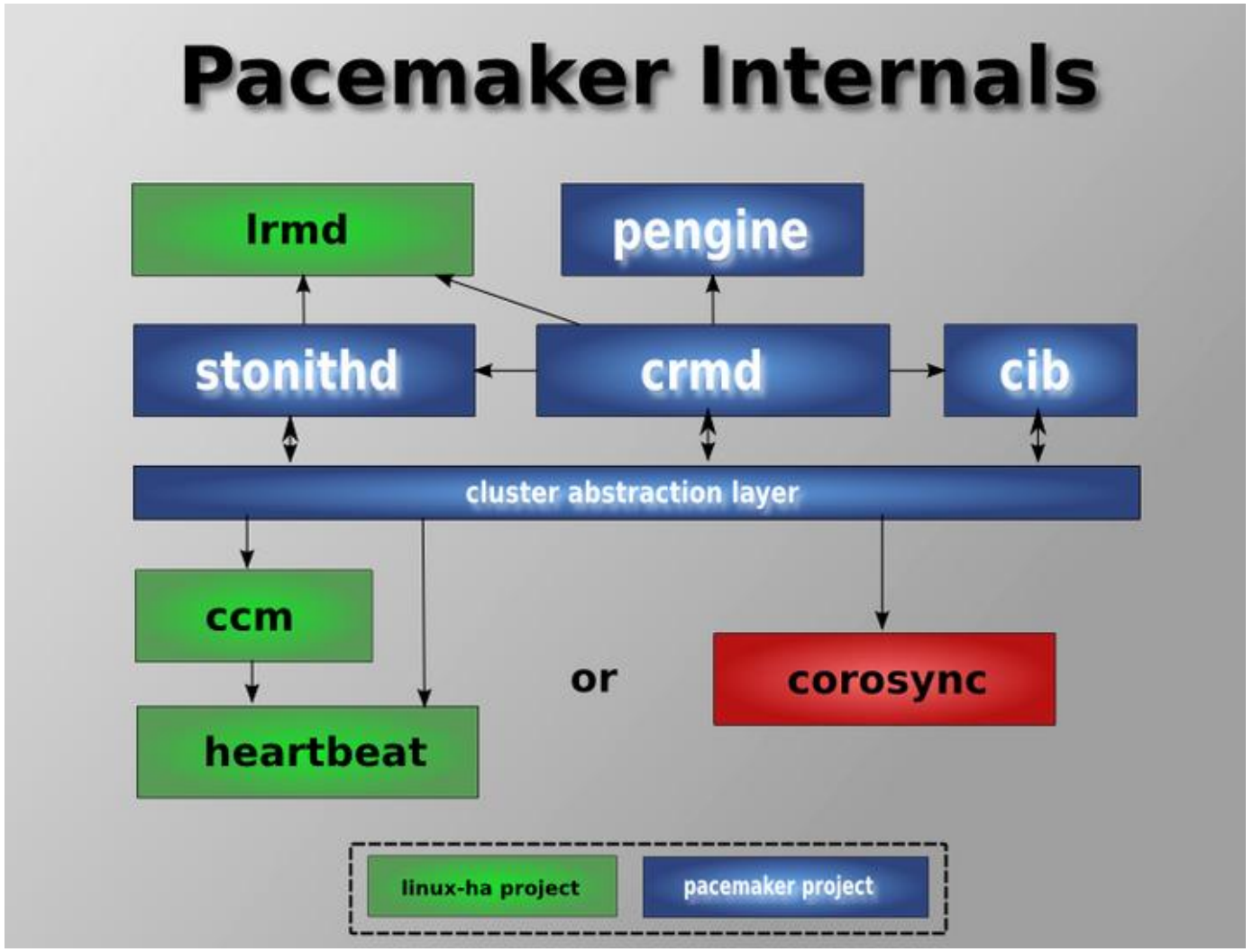




两台设备的stonith连接示例

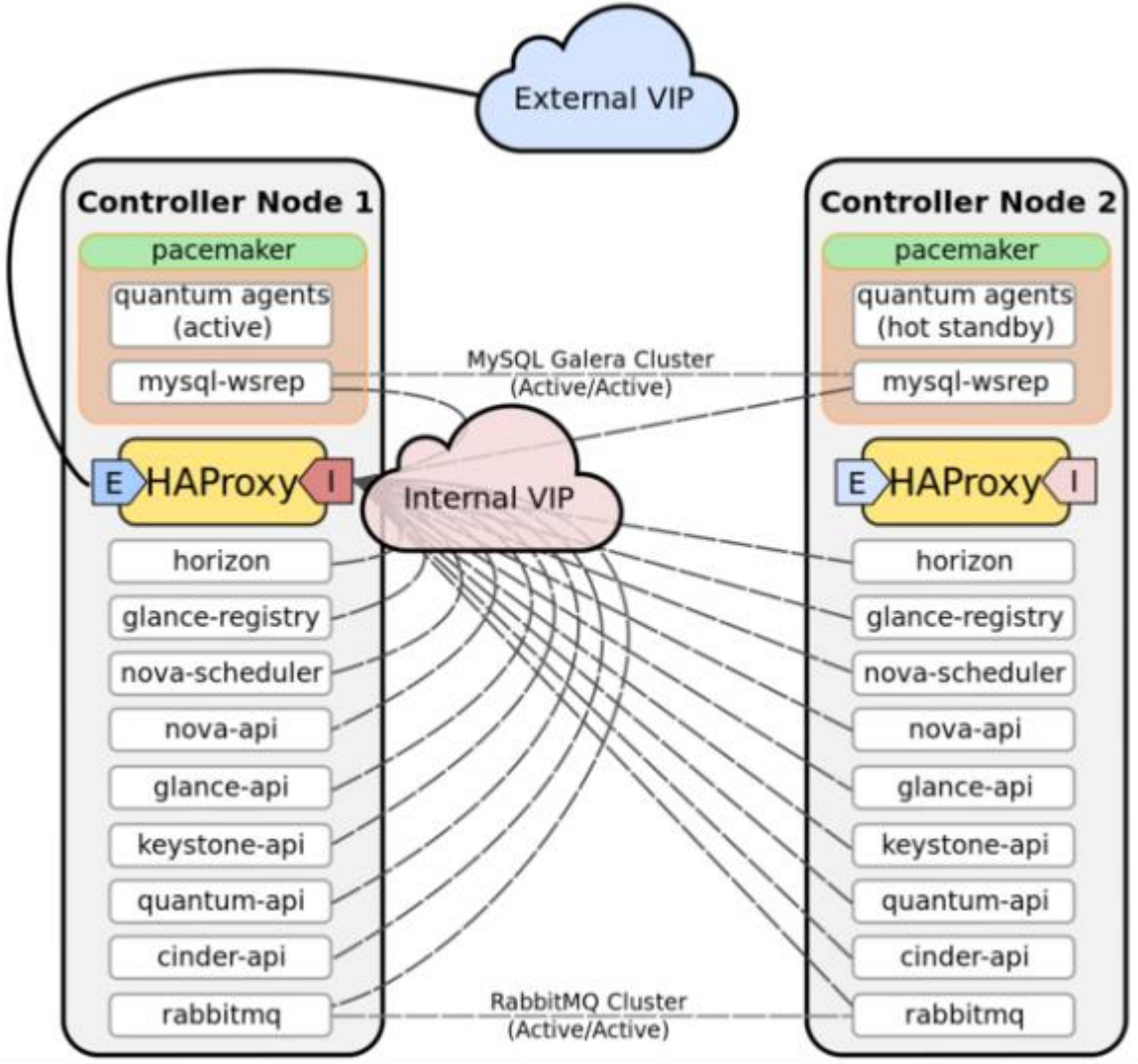
stonith是“shoot the other node in the head”的首字母简写，它是Heartbeat软件包的一个组件，它允许使用一个远程或“智能的”连接到健康服务器的电源设备自动重启失效服务器的电源

Pacemaker入门



Pacemaker本身由四个关键组件组成:
•CIB (集群信息基础)
•CRMd (集群资源管理守护进程)
•PEngine (PE or 策略引擎)
•STONITHd，STONITH设备被当成资源，配置在CIB中配，从而轻松地监控

Corosync +Pacemaker案例



Thanks

FAQ时间