

搭建基于 Linux 具有高可用性的集群环境

反馈：史应生 shiyingsheng@yahoo.com

高可用性是企业级服务器集群的一个重要元素，可以帮助在服务器宕机的情况下减小服务的 "downtime".

本文章从技术的角度，讲述了目前主流的 Linux 发行版 (Red Hat 和 Novell) 的高可用产品的构架和特性。

适合读者：

中/高级的 Linux 系统管理员
企业 IT 部门的决策者
方案构架师
对高可用有兴趣的所有人

一. 高可用概念：

高可用集群软件通常要包括几个通用的特性。至少要提供：

1. 一种机制来定义哪些系统可以被用作集群节点
2. 哪些服务或者应用可以在节点间作失效切换 (fail-over)
3. 节点间内部相互通信的方式
4. 当失效的节点控制相同的集群资源的情况下,防止资源的冲突
5. 防止集群裂脑 (split-brain) 发生
6. Fence 机制或者更加复杂的 I/O fence 机制
7. 提供集群合作管理的机制
8. 提供监控工具
9. 预先定义的应用和服务的监控脚本

二. Heartbeat 和 SUSE Linux Enterprise Server

Heartbeat 来自于 High-Availability 项目 (www.linux-ha.org)。SLES9 和 SLES10 所带的版本不同，SLES9 包含的版本是 Heartbeat1.x，它允许创建 2 个节点的集群，提供基本的高可用性 failover 服务。SLES10 包含的版本是 Heartbeat2.x。它允许创建多个节点的集群，提供增强的特性

2.1) Heartbeat 1.x 的特性

Heartbeat1.x 允许集群节点和资源通过 /etc/ha.d 目录下面的两个文件来配置

ha.cf：定义集群节点，失效检测和切换时间间隔，集群时间日志机制和节点 Fence 方法

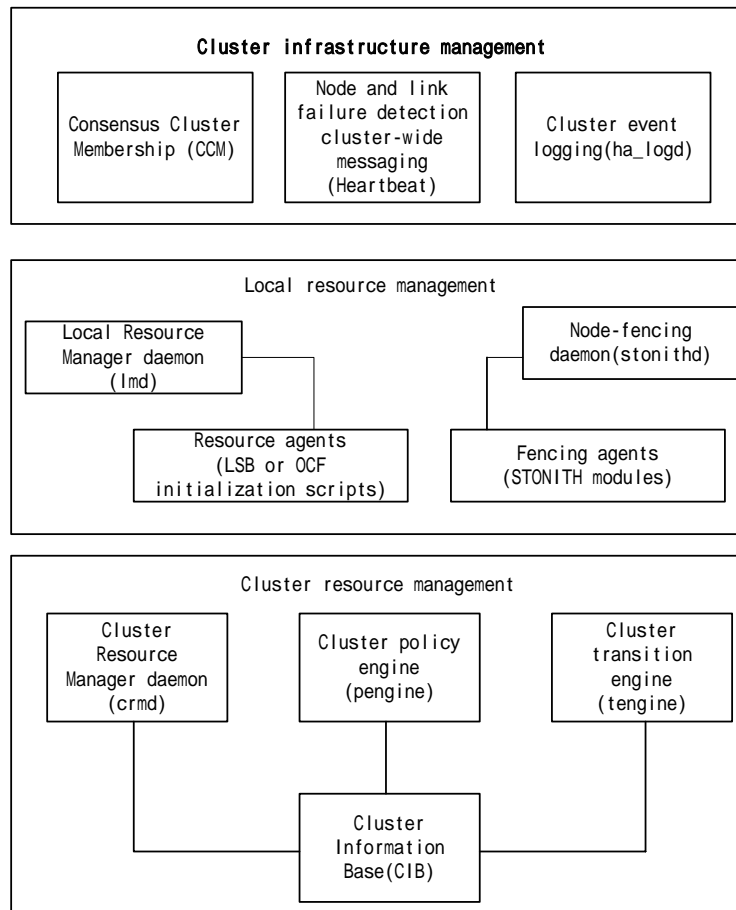
haresources：定义集群资源组，每一行定义可以一起进行失效切换的一个默认的节点和一组资源，资源包括 IP 地址，文件系统，服务或者应用

2.2) Heartbeat 2.0 的特性

Heartbeat 2.0 即支持基于 Heartbeat1.x 的配置（仅限于 2 个节点）又支持模块结构的配置方法 - 集群资源管理器(Cluster Resource Manager-CRM).

CRM 模型可以支持最多 16 个节点，这个模型使用基于 XML 的集群信息(Cluster Information Base-CIB)配置。CIB 文件(/var/lib/heartbeat/crm/cib.xml)会在各个节点间自动复制，它定义了下面的对象和动作：

- *集群节点
- *集群资源，包括属性，优先级，组和依赖性
- *日志，监控，仲裁和 fence 标准
- *当服务失败或者其中设定的标准满足时，需要执行的动作



图一

图一显示了 Heartbeat2.0 结构的关键元素。

Consensus Cluster Membership 服务使用选举机制允许集群节点决定指定的协调器 (Designated Coordinator - DC)，它来帮助建立仲裁，管理集群节点成员关系和资源分配。DC 维护集群的状态和管理策略。其他的节点必须转发状态改变请求到 DC 处理。Heartbeat 服务检查节点和连接状态来决定失效是否发生，集群事件日志服务 (ha-logd) 提供集群套件中所有服务的日志功能。

为了控制集群资源，本地资源管理器(Local Resource Manager-LRM)启动，停止和监控资源代理。LRM 守护进程(lmd)负责和 DC 的集群时间通信。节点的 fence 代理是一种特殊的资源，由 node-fencing 进程 stonithd 控制。stonithd 的意思是 "Shoot the Other Node in the Head",

主要是使出现问题的节点从集群环境中脱离。fence 设备包括串行或者基于网络的电源切换设备或者远程管理硬件。

当节点不能正常通信时，fence 防止不同子集的节点运行相同的资源。这种情况叫做裂脑。裂脑通过使用应用设计，节点 fencing 或者资源指定的 fencing 来避免。

CRM 守护进程(crmd)管理 CIB,它允许对节点和资源的行为的高级限制和依赖。集群策略引擎 (pengine)解释和实施这些限制和依赖。集群转移引擎(tengine)管理 CRM 的状态和在出现失效事件时协调在另一个节点上进程的重新启动和资源转移。

2.3) 配置工具

Heartbeat2.0.5,包含在 SLES10 中，引入了 GUI 工具用于集群的管理和监控。它包括监控脚本样本来协助通用 Linux 服务和应用的配置，包括基于 xinetd 的服务，Apache 服务，IBM DB2 数据库，IBM WebSphere 应用服务器。许多其他的应用，例如 NFS,Samba 也可以进行配置。

Heartbeat2.0 遵循 Open Cluster Framework (OCF)资源代理应用编程接口，允许使用通用的 LSB 初始化脚本和集群相关的 OCF 资源初始化脚本。

每个版本的 Heartbeat 也可以配置为结合 Linux Virtual Server 功能的用于 IP 负载均衡的功能。这取决于被配置的服务和资源的需求，共享存储，是否使用集群文件系统的并发访问机制等等，与 Oracle 的 OCFS 的结合会在 Heartbeat 的下一个版本中集成。

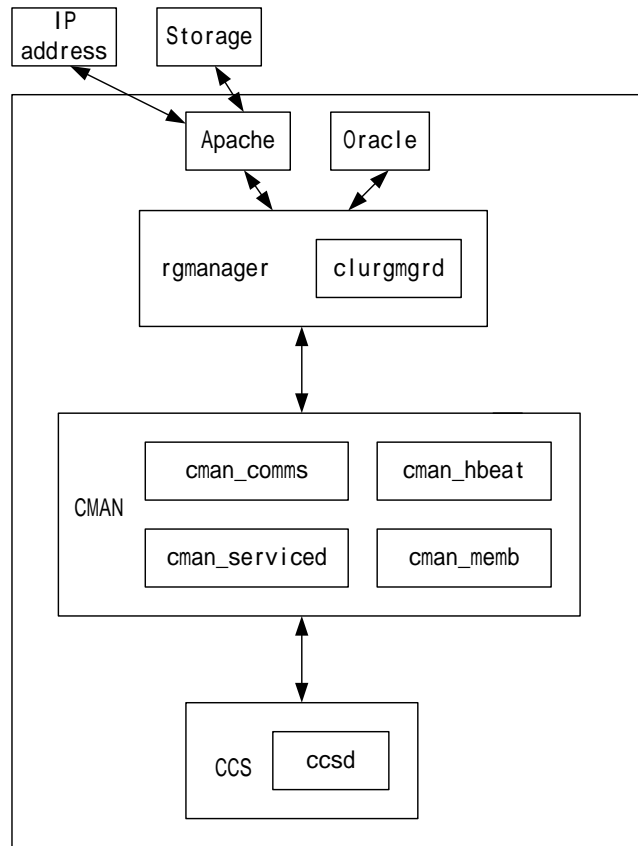
在 SLES10 中包含在 Heartbeat2.0.5 中的 GUI 工具简化了配置。同时，Novell 也计划使用 Heartbeat2.0 的核心服务作为将来 Novell Cluster Services(NCS)软件的基础。NCS 是独立 license 的，包括了预先定义的资源类型,GUI 配置工具和监控工具。

三. Red Hat Cluster Suite (RHCS) 和 Red Hat Enterprise Linux 4

RHCS 专门为 Red Hat Enterprise Linux 设计包含了下面两个不同类型的集群

应用和服务切换：创建关键应用和服务的多节点服务器的集群环境

IP 负载均衡：对于进来的 IP 网络请求在一群服务器组中做负载均衡



图二

集群的主要元素包括 Cluster Manager (CMAN), Cluster Configuration System (CCS) 和 Resource Group Manger (rgmanager). 图二显示了在任何指定的时间运行在一个节点上的不同的服务和守护进程的关系

CCS 提供访问位于每一个节点的单一集群配置文件/etc/cluster/cluster.conf. 配置文件包括版本号, 它在集群任何时候改变时都会更新. ccsd 运行在每一个节点上. 当 ccsd 启动后, 它找到节点间最新版本的配置文件.

CMAN 用于管理集群成员, 消息和通知. CMAN 包括一套内核补丁和一个用户空间程序 (cman_tool).

cman_tool 用于使一个节点加入或者离开集群. 改变集群的投票期望值. CMAN 依赖于 CCS.

组资源管理器进程 (clurgmgrd) 处理管理员指定的集群服务 (也称之为资源), 包括管理员的请求比如服务启动, 服务禁止, 服务重新加载和服务重启动. 它也处理在服务失效时, 服务的重新启动和服务重定向.

3.1 配置工具

RHCS 支持 16 个节点的集群. GUI 的配置工具是 system-config-cluster. 集群配置包括: 资源信息, 节点信息, fencing 设备信息和失效域信息. 这些信息以 XML 的格式存储在每个节点的 /etc/cluster/cluster.conf 文件中. 这些资源在一个服务下被组织成资源组.

失效域是集群成员的子集。失效域有一下的特性：

无限制 — 允许你指定要优选的成员子集，但是被分派到这个域的服务可以在任何可用的成员上运行。

有限制 — 允许你限制能够运行某个特定服务的成员。如果在限制的失效转移域中没有一个可用的成员，服务就无法被启动（手工启动或被群集软件启动）。

无序 — 当服务被分派给一个无序的失效转移域，运行服务的成员就会从失效转移域成员中不按优先顺序被选择。

有序 — 允许你在失效转移域成员中指定一个优先顺序。在列表最前面的是最优先的，跟着是次一级的，依此类推。

按照默认设置，失效转移域是无限制和无序的。

CCS 支持集群信息的在线改变，而且会自动同步到其他的节点。

3.2 失效切换能力

类似于 STONITH, fence 设备是一个节点在它重新启动它的服务前可以 power cycle 另一个节点。

Fence 设备可以在一个没有响应的节点恢复后，防止数据冲突。如果 CMAN 检测到一个节点失败，失败的节点会从集群中删除。如果不使用 fence 设备，那么一个失效的节点可能会导致集群服务在多余一个节点运行，从而造成数据冲突甚至是系统崩溃。

四．针对这两个厂家的商业高可用性解决方案如何选择：

表 1 对 Red Hat 和 Novell 高可用软件的作了技术比较。

特点	SUSE Linux Enterprise Server with Heartbeat	Red Hat Cluster Suite
部署,配置和管理	GUI 工具简化了配置和管理工作	GUI 工具简化了配置和管理工作
如何得到软件	包含在 SLES 中,或者 www.linux-ha.org	单独购买
最大节点数支持	Heartbeat1.x 支持 2 个节点 Heartbeat2.x 支持 16 个节点	16 个节点
资源类型	IP 地址,文件系统,NFS 输出,Samba, Apache 和 LSB 与 OCF 初始化脚本	IP 地址,文件系统,NFS 输出,Samba, 初始化脚本,红帽 GFS
Fencing 设备	支持各种串行和网络电源切换器,远程管理卡	支持各种串行和网络电源切换器,远程管理卡
共享存储	SAN,NAS,ISCSI (单机文件系统/集群文件系统)	SAN,NAS,ISCSI (单机文件系统/集群文件系统)

表一

个人认为：

对于费用敏感的企业 Heartbeat 是个不错的选择，它在 Novell 的发行版中提供。

对于想寻求易用的企业可以选在 RHCS 和它的 GUI 功能。RHCS 和 RHEL 是独立的产品