Red Hat集群套件介绍

Red Hat Cluster Suite 是一款能够提供高性能、高可靠性、负载均衡、高可用性且经济廉价的集群工具集。

一个集群通常有两个或两个以上的计算机（称为“节点”或”成员“）共同执行一个任务。其中有四种常见集群：

\* 存储集群

\* 高可用性集群

\* 负载均衡

\* 高性能集群

存储集群在一个集群中为服务提供一个一致的文件系统映像，允许服务同时去读写一个单一的共享文件系统。存储集群通过将数据放到一个共享文件系统中从而消除了在应用程序间拷贝数据的麻烦，并提供一个单一的备份和故障恢复点。Red Hat Cluster Suite通过Red Hat GFS提供一个存储集群。

高可用性集群通过消除单一故障点和节点故障转移功能（当一个集群节点失败后将服务转移到其他节点上）来提供高可用性。节点故障转移功能对客户端是透明的，当节点失败后客户端并不会看到节点之间的服务转移。Red Hat Cluster Suite通过高可用性服务管理组件来提供一个高可用性集群。

负载均衡集群将服务请求调度到集群中的多个节点上。负载均衡是一个低成本、高可用性的集群，因为你可以根据负载情况灵活的添加和删除节点。在负载均衡集群中当一个节点失败后，调度器会发现这个失败并停止向此节点发送请求。节点的失败对于客户端是透明的。Red Hat Cluster Suite 通过LVS（Linux Virtual Server）来提供负载均衡集群。

高性能集群在节点上执行并行计算。高性能集群允许应用程序并行计算从而提高应用程序性能。

RHCS主要有下面部分组成：

\* 集群架构--提供一个基本功能使节点作为集群工作在一起：配置文件管理，成员关系管理，锁管理和栅设备。

\* 高可用性服务管理--提供节点失败转移服务，当一个节点失败后将服务转移到另一个节点上。

\* 集群管理工具--通过配置和管理工具来配置和管理Red Hat集群。

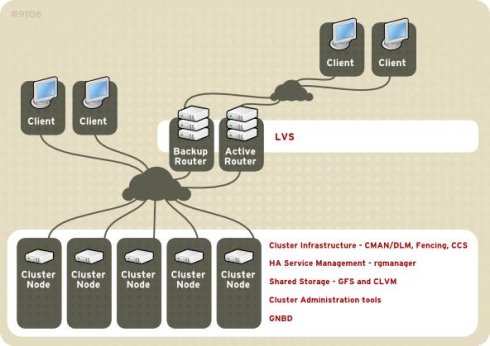
\* Linux Virtual Server (LVS)--LVS提供一个基于IP的负载均衡功能，通过LVS可以将客户请求均匀的分配到集群节点上。

你可以通过下面的组件补充Red Hat集群：

\* Red Hat GFS (Global File System)--GFS为Red Hat Cluster Suite提供一个集群文件系统，CFS允许多个节点在块级别上共享存储。

\* Cluster Logical Volume Manager (CLVM)--提供逻辑卷管理集群存储。

\* Global Network Block Device (GNBD)--GFS的一个补充组件用于将存储在块级别导出到以太网上。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static5.photo.sina.com.cn/orignal/4e424e21441641944ea04)

Red Hat Cluster Suite 介绍

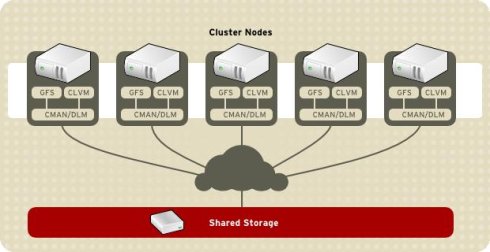
**Red Hat集群介绍**

**1、集群管理**

CMAN（cluster manager）在Red Hat Cluster Suite中执行集群管理任务。CMAN是一个分布式集群管理器，它运行在每一个集群节点上。

CMAN通过监视集群节点提供一个法定节点数（quorum），当集群中有多余一半的节点处于活跃状态时说明有一个法定节点数，当有一半或少于一半的节点处于活跃状态是则不具有一个法定节点数，此时整个集群变得不可用。

CMAN通过监控节点来了解节点的成员关系，当集群中的成员关系改变是，CMAN通知底层结构做相应调整。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static5.photo.sina.com.cn/orignal/4e424e21441641959e584)

CMAN/DLM 概览

**2、锁管理**

锁管理是一个公共的基础结构，对其它集群结构提供一种共享集群资源的机制。在Red Hat集群中DLM是一个分布式锁管理器，它运行在所有的集群节点上。GFS和CLVM通过锁管理器来使用锁。GFS通过锁管理器使用锁机制来同步访问文件系统元数据。CLVM通过锁管理器来同步更新数据到LVM卷和卷组。

**3、栅设备（Fencing）**

通过栅设备可以从集群共享存储中断开一个节点。栅设备从共享存储切断I/O以保证数据的完整性。当CMAN确定一个节点失败后，它在集群结构中通告这个失败的节点，fenced进程将失败的节点隔离，以保证失败节点不破坏共享数据。Fenced运行在每个节点上。

Red Hat Cluster Suite提供下面集中隔离方式：

\* 电源栅设备

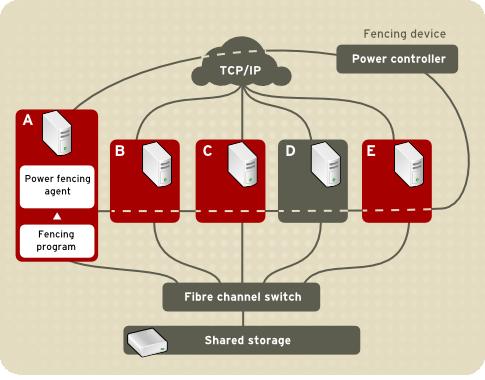
\* 光线通道栅设备

\* GNBD栅

\* 其他栅设备

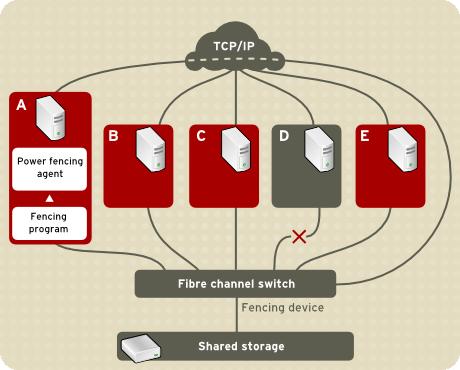
栅设备实例

当A上的栅进程发现D节点失效时，它通过栅代理通知电源控制器将D节点隔离。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static6.photo.sina.com.cn/orignal/4e424e214416419563865)

电源栅设备实例

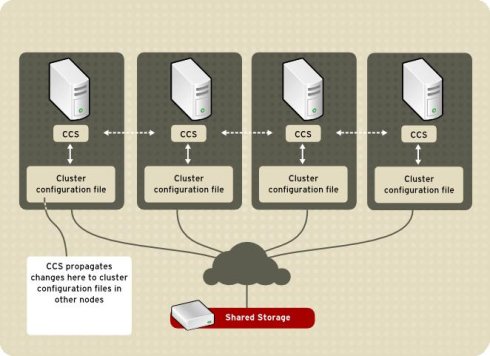
当A上的栅进程发现D节点失效时，它通过栅代理通知光纤通道交换设备将D节点隔离。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static7.photo.sina.com.cn/orignal/4e424e21441641959fcc6)

光纤通道交换栅设备实例

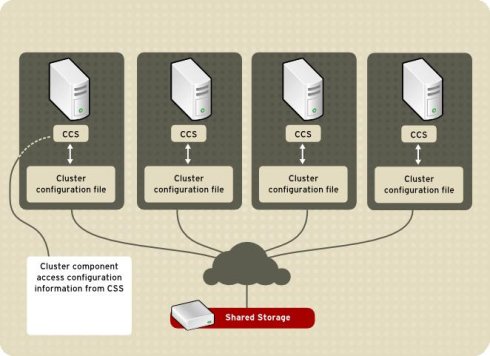
**4、集群配置系统**

集群配置系统（CCS）管理集群配置，并为其他集群结构提供配置信息。CCS运行在每个节点上并确保集群配置文件在每个节点上是最新的。举例，假如管理员在节点A上更新配置文件，CCS负责传播更新到其他节点上。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static10.photo.sina.com.cn/orignal/4e424e21441641961f329)

CCS概览

其他集群组件（如CMAN）通过CCS从配置文件访问配置信息。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static13.photo.sina.com.cn/orignal/4e424e21441641968718c)

访问配置信息

集群配置文件(/etc/cluster/cluster.conf) 是一个XML文件，用来描述下面的集群特性：

\* 集群名称--列出集群名称、集群配置文件版本和一个隔离时间，隔离相应时间当一个新节点加入或从集群中隔离时。

\* 集群--列出集群中的每一个节点，指定节点名称，节点ID，法定投票数，和栅模式。

\* 栅设备--定义栅设备。

\* 管理资源--定义创建集群服务需要的资源。管理资源包括失败转移域，资源（如IP）和服务。

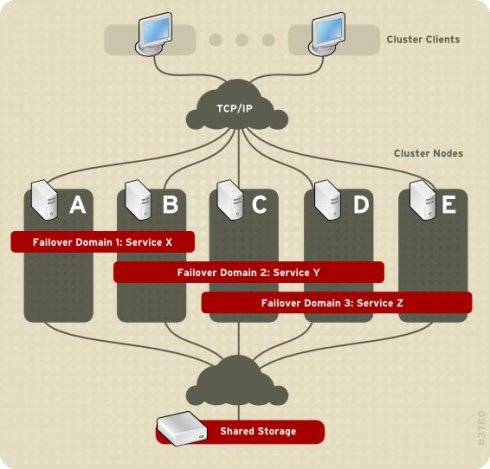
**5、高可用性服务管理**

高可用性服务管理提供一种创建高可用性集群服务的能力。rgmanager运行在每个集群节点上。当一个节点失败时高可用性集群能将服务从一个失败节点转移到其他节点上，而这种转移对客户端是透明的。

你可以在集群配置文件中配置一个高可用性集群服务。集群服务包含集群资源（如一个IP地址，一个应用程序初始化脚本，Red Hat GFS共享文件系统等）。

你可以将一个集群服务和一个失败转移域相结合。一个失败转移域是一个运行特定服务的集群节点子集。在一个失败转移域中可以为每个节点指定一个优先级，来决定在节点发生失败时服务转移顺序。假如你未指定一个优先级，集群服务可以在失败转移域中的任意节点间转移。你也可以限制一个服务只在失败转移域间转移。

下面是一个例子，失败转移域1由节点A和B组成，并配置了限制，当节点A失败后，服务X只能转移到节点B上。失败转移域2由节点B、C和D组成，并配置了限制，C的优先级设置为1，B的优先级设置为2，D的优先级设置为3，当C节点失败时，服务Y试图转移到B，如果不成功则转移到D。失败转移域3配置为无优先级、无限制，如果服务Z所在节点失败它将试图转移到其它节点，当域中的节点均无效时服务转移到集群中的其它节点。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static9.photo.sina.com.cn/orignal/4e424e214416419662578)

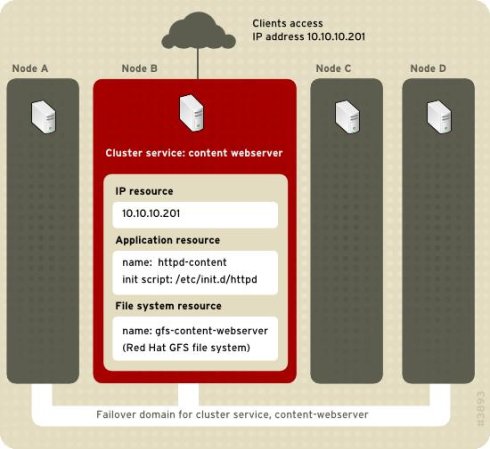
失败转移域

下面是一个高可用性集群的例子，有一个叫做"content-webserver".的服务运行在节点B上，有一个失败转移域包括节点A、B和D。另外失败转移域配置一个优先级，D优先级高于A，并设置限制。这个服务包含下面集群资源：

\* IP地址资源--IP地址10.10.10.201。

\* 应用程序资源"httpd-content"--一个Web服务程序初始化脚本/etc/init.d/httpd 。

\* 文件系统资源--GFS名字为："gfs-content-webserver"。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static4.photo.sina.com.cn/orignal/4e424e214416419736c13)

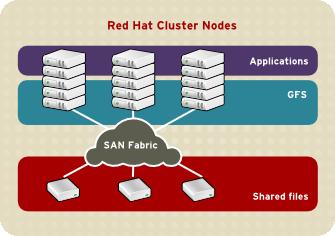
Web Server Cluster Service示例

客户端通过10.10.10.201访问web服务，httpd-content 应用程序将数据放到gfs-content-webserver文件系统中。

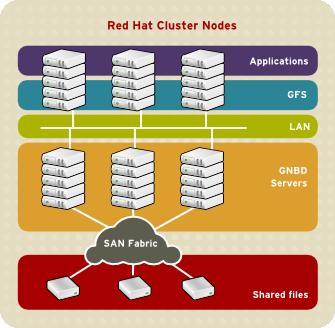
**6、Red Hat GFS**

GFS为Red Hat Cluster Suite提供一个集群文件系统，它允许多个节点在块级别上共享存储。GFS使用一个分布式的元数据和多个订阅数（multiple journals），GFS是一个理想的集群文件系统。为了保证文件系统的完整性，GFS使用锁管理器来管理I/O。当一个节点在GFS上修改数据时，改变可以立即在其它使用此文件系统的节点上可见。

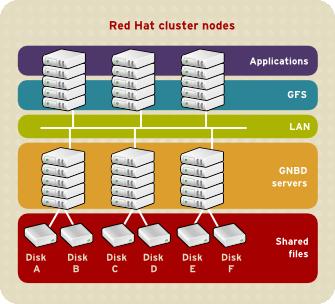
运行Red Hat GFS的节点可以通过Red Hat Cluster Suite配置管理工具进行管理。通过CLVM来进行卷管理。GFS在集群节点之间提供数据共享。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static7.photo.sina.com.cn/orignal/4e424e21441643195f5b6)

GFS with a SAN

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static1.photo.sina.com.cn/orignal/4e424e21441643197a0a0)

GFS and GNBD with a SAN

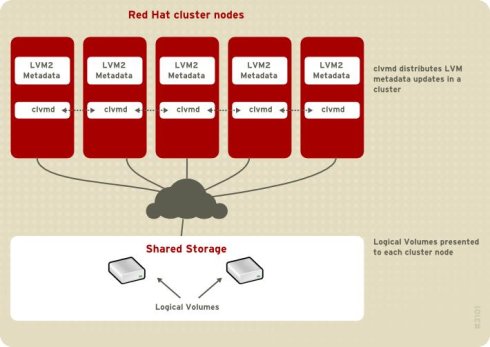
[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static13.photo.sina.com.cn/orignal/4e424e214416431ad9a2c)

GFS and GNBD with Directly Connected Storage

**7、集群逻辑卷管理器**

CLVM在一个单一节点上提供与LVM2一样的性能，但它可以创建可被所有集群节点可用的卷。通过CLVM创建的逻辑卷可以被用于所有的集群节点上。

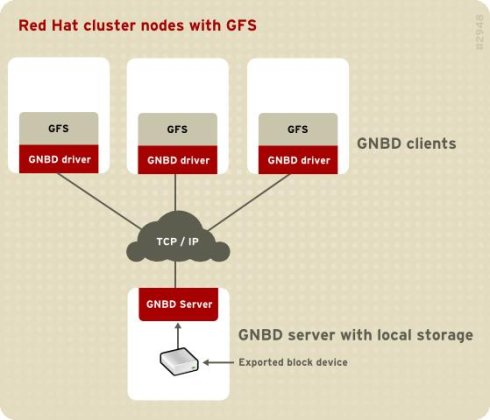
clvmd对LVM2标准工具提供一个集群扩展，使其允许使用LVM2工具管理共享存储。clvmd运行在每个集群节点上，在集群中分布式更新元数据从而可以是其它节点看到一个一致的逻辑卷视图。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static9.photo.sina.com.cn/orignal/4e424e214416431a8c378)

CLVM 概览

**8、Global Network Block Device（GNBD)**

Global Network Block Device (GNBD) 提供通过TCP/IP访问Red Hat GFS的方法。GNBD由两个主要部分组成：一个GNBD客户端，一个GNBD服务器端。GNBD服务器用于导出本地存储。GNBD客户端运行在使用GFS的节点上，导入从GNBD服务器端导出的磁盘设备。多个客户端可以同时访问有一个GNBD服务器导出的磁盘设备。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static16.photo.sina.com.cn/orignal/4e424e214416431a6f19f)

GNBD概览

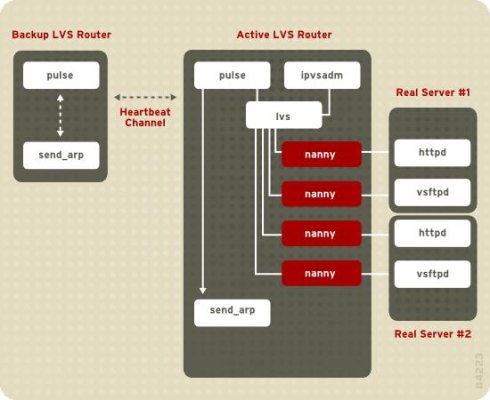
**9、Linux Virtual Server**

Linux Virtual Server (LVS)是一套完整的基于IP的负载均衡的集群软件。LVS运行在一对有相似配置的计算机上：一个作为活动LVS Router，一个作为备份LVS Router。活动LVS Router服务有两个角色：

\* 均衡负载到真实服务器上。

\* 检查真实服务器提供的服务是否正常。

备份LVS Router用来监控活动的LVS Router，以备活动的LVS Router失败时由备份LVS Router接管。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static6.photo.sina.com.cn/orignal/4e424e214416431b43b55)

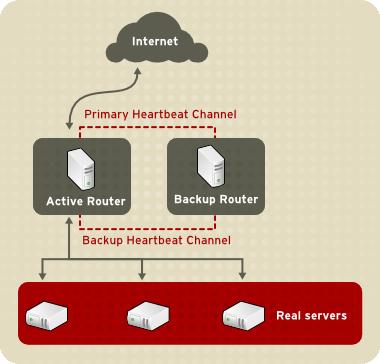
LVS Cluster

Pulse进程运行在活动LVS Router和备份LVS Router上。在备份LVS Router上，pulse发送一个心跳（heartbeat）到活动LVS Router的公网接口上以检查活动LVS Router是否正常。在活动LVS Router上，pulse启动lvs进程并响应来自于备份LVS Router的心跳。

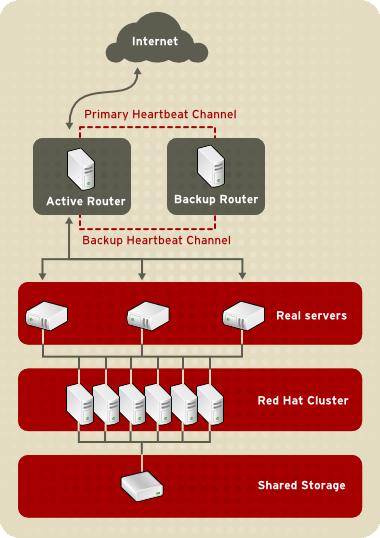
lvs进程调用ipvsadm工具去配置和维护IPVS路由表，并为每一个在真实服务器上的虚拟服务启动一个nanny进程。每一个nanny进程去检查真实服务器上的虚拟服务状态，并将故障情况通知lvs进程。假如一个故障被发现，lvs进程通知 ipcsadm在IPVS路由表中将此服务器删除。

如果备份LVS Router未收到来自于活动LVS Router的响应，它将调用send\_arp将虚拟IP地址再分配到备份LVS Router的公网接口上。并在公网接口和局域网接口上分别发送一个命令去关掉活动LVS Router上的lvs进程。同时启动自己的lvs进程来调度客户端请求。

下面是一个两层结构的LVS架构：

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static12.photo.sina.com.cn/orignal/4e424e214416431b276db)

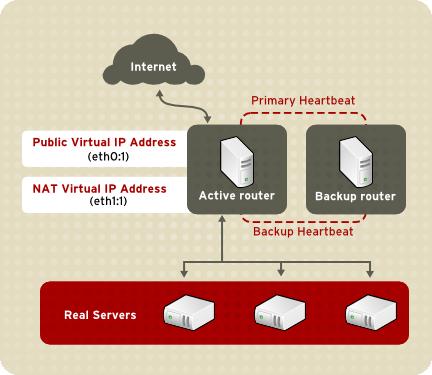
下面是一个三层结构的LVS架构：

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static16.photo.sina.com.cn/orignal/4e424e214416431c53abf)

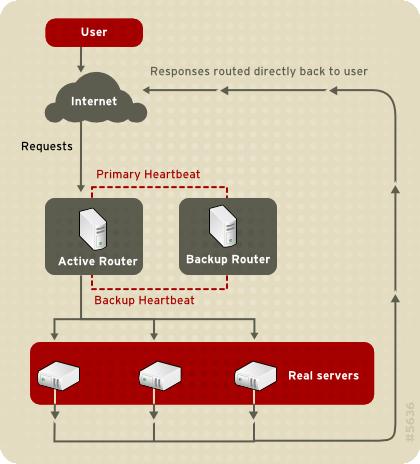
两层结构的LVS架构适合于Real Server上的数据变化不是很频繁的情况，如果Real Server上的数据变化很频繁应该考虑使用三层结构的LVS架构，将数据放到一个共享文件系统（GFS）上供所有的Real Server同时读写。

你可以使用VS/NAT和VS/DR来实现IP负载均衡：

\* 使用VS/NAT方法：客户通过Virtual IP Address（虚拟服务的IP地址）访问网络服务时，请求报文到达调度器，调度器根据连接调度算法从一组真实服务器中选出一台服务器，将报文的目标地址 Virtual IP Address改写成选定服务器的地址，报文的目标端口改写成选定服务器的相应端口，最后将修改后的报文发送给选出的服务器。同时，调度器在连接Hash 表中记录这个连接，当这个连接的下一个报文到达时，从连接Hash表中可以得到原选定服务器的地址和端口，进行同样的改写操作，并将报文传给原选定的服务器。当来自真实服务器的响应报文经过调度器时，调度器将报文的源地址和源端口改为Virtual IP Address和相应的端口，再把报文发给用户。当使用VS/NAT方法时，如果有大量的响应数据经过调度器，调度器将成为整个集群的瓶颈。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static10.photo.sina.com.cn/orignal/4e424e21441643ca4b019)

\* 使用VS/DR方法：调度器和服务器组都必须在物理上有一个网卡通过不分断的局域网相连，如通过交换机或者高速的HUB相连。VIP地址为调度器和服务器组共享，调度器配置的VIP地址是对外可见的，用于接收虚拟服务的请求报文；所有的服务器把VIP地址配置在各自的Non-ARP网络设备上，它对外面是不可见的，只是用于处理目标地址为VIP的网络请求。在VS/DR中，调度器根据各个服务器的负载情况，动态地选择一台服务器，不修改也不封装IP报文，而是将数据帧的MAC地址改为选出服务器的MAC地址，再将修改后的数据帧在与服务器组的局域网上发送。因为数据帧的MAC地址是选出的服务器，所以服务器肯定可以收到这个数据帧，从中可以获得该IP报文。当服务器发现报文的目标地址VIP是在本地的网络设备上，服务器处理这个报文，然后根据路由表将响应报文直接返回给客户。

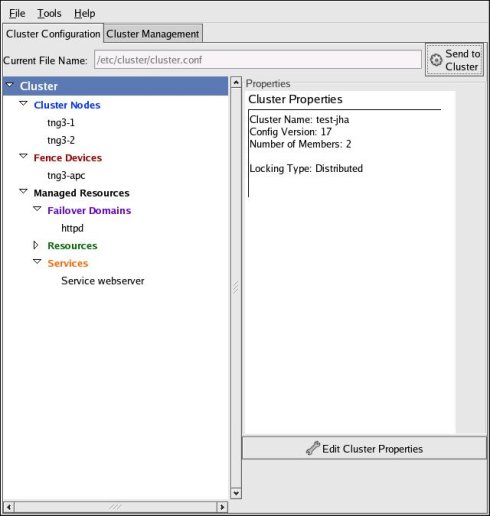
[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static5.photo.sina.com.cn/orignal/4e424e21441643cc2c734)

**10、集群管理工具**

Red Hat Cluster Suite提供了多种集群配置和管理工具，如Conga，Cluster Administration GUI和基于命令行的管理工具。

Conga是一款基于Web的管理工具，它提供一个集中管理Red Hat集群和存储的管理平台。它主要有luci和ricci两部分组成，其中ricci运行在每一个集群节点上，luci安装在一台用于管理集群的计算机上。Luci通过ricci和集群中的节点通讯。

system-config-cluster 也是一款用于配置和管理集群节点的图形化管理工具。它有集群配置和集群管理两个选项卡组成，用于完成相应的配置和管理工作。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static16.photo.sina.com.cn/orignal/4e424e21441643cd1c18f)

ccs\_tool、cman\_tool、fence\_tool、clustat、clusvcadm是几款命令行工具，具体请看相应的man页。

Piranha是一款基于web页面的用于配置和管理LVS集群的图形工具。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=4e424e2101007qpm&url=http://static7.photo.sina.com.cn/orignal/4e424e21441643ccde266)