**【教程主题】：3.LVS**

**【课程录制】： 创E**

**【主要内容】**

**【1.1】集群简介**

集群的概念(Concepts of Cluster)

HAC(High Availability Cluster)

LBC(Load Balance Cluster)

HPC(High Performance Computing Cluster)

Grid computing

计算机集群简称集群是一种计算机系统， 它通过一组松散集成的计算机软件和/或硬件连接起来高度紧密地协作完成计算工作。在某种意义上，他们可以被看作是一台计算机。集群系统中的单个计算机通常称为节点，通常通过局域网连接，但也有其它的可能连接方式。集群计算机通常用来改进单个计算机的计算速度和/或可靠性。一般情况下集群计算机比单个计算机，比如工作站或超级计算机性能价格比要高得多。

集群分为同构与异构两种，它们的区别在于：组成集群系统的计算机之间的体系结构是否相同。集群计算机按功能和结构可以分成以下几类:

高可用性集群 High-availability (HA) clusters

负载均衡集群 Load balancing clusters

高性能计算集群 High-performance (HPC) clusters

网格计算 Grid computing

**高可用性集群**

当集群中有某个节点失效的情况下，其上的任务会自动转移到其他正常的节点上。可以将集群中的某节点进行离线维护再上线，该过程并不影响整个集群的运行。

**负载均衡集群**

负载均衡集群运行时，一般通过一个或者多个前端负载均衡器，将工作负载分发到后端的一组服务器上，从而达到整个系统的高性能和高可用性。这样的计算机集群有时也被称为服务器群（Server Farm）。 一般高可用性集群和负载均衡集群会使用类似的技术，或同时具有高可用性与负载均衡的特点。

Linux虚拟服务器（LVS）项目在Linux操作系统上提供了最常用的负载均衡软件。

**高性能计算集群**

高性能计算集群采用将计算任务分配到集群的不同计算节点而提高计算能力，因而主要应用在科学计算领域。比较流行的HPC采用Linux操作系统和其它一些免费软件来完成并行运算。这一集群配置通常被称为Beowulf集群。这类集群通常运行特定的程序以发挥HPC cluster的并行能力。这类程序一般应用特定的运行库, 比如专为科学计算设计的MPI库。

HPC集群特别适合于在计算中各计算节点之间发生大量数据通讯的计算作业，比如一个节点的中间结果或影响到其它节点计算结果的情况。

**网格计算**

网格计算或网格集群是一种与集群计算非常相关的技术。网格与传统集群的主要差别是网格是连接一组相关并不信任的计算机，它的运作更像一个计算公共设施而不是一个独立的计算机。还有，网格通常比集群支持更多不同类型的计算机集合。

网格计算是针对有许多独立作业的工作任务作优化，在计算过程中作业间无需共享数据。网格主要服务于管理在独立执行工作的计算机间的作业分配。资源如存储可以被所有结点共享，但作业的中间结果不会影响在其他网格结点上作业的进展。

**集群技术特点**

1是通过多台计算机完成同一个工作。达到更高的效率。 2是两机或多机内容、工作过程等完全一样。如果一台死机，另一台可以起作用。

**【1.2】LVS历史概述**

Linux Virtual Server项目的目标：使用集群技术和Linux操作系统实现一个高性能、高可用的服务器，它具有很好的可伸缩性（Scalability）、可靠性（Reliability）和可管理性（Manageability）。在某种程度上是对市场上商业化硬件集群解决方案的开源替代方案。

LVS项目创始人章文嵩，1973年生人，国防科技大学博士。曾与LIDS项目创始人谢华刚联合创建中国集群技术有限公司。2009年9月加盟淘宝网，为淘宝资深总监、研究员、淘宝技术委员会主席，主要负责淘宝网基础核心软件研发、推进网络软硬件方面的性能优化、搭建下一代高可扩展低碳低成本淘宝电子商务基础设施。

在1998年5月，章文嵩先生成立了Linux Virtual Server的自由软件项目，进行Linux服务器集群的开发工作。同时，Linux Virtual Server项目是国内最早出现的自由软件项目之一。LVS第一个版本只花了两个星期的时间便把开源软件代码完成了，然后放到网上。因为是开源的，到网站上下载的人很多。LVS项目于1998年5月在网站上发布IPVS第一个版本源程序以来，一直得到了来自 Internet 的用户和开发者的鼓励和支持。

2004年12月24日，ipvs-1.2.1作为Netfilter的模块正式入驻linux官方内核2.6.10，得到更广泛地传播和应用。

官方网站：<http://www.linuxvirtualserver.org/>

**【1.3】LVS介绍**

基于LVS的负载均衡集群

三种网络/工作模式：NAT、DR、TUN

十个调度算法： rr、wrr、lc、wlc、lblc、lblcr、dh、sh、sed、 nq

集群系统的角色、结构

Client: CIP

Router:GWIP

VirtualServer(VS): VIP

Director/LoadBalancer(LB): DIP

RealServer(RS): RIP

1.轮叫调度（Round Robin）(简称rr)

调度器通过“轮叫”调度算法将外部请求按顺序轮流分配到集群中的真实服务器上，它均等地对待每一台服务器，而不管服务器上实际的连接数和系统负载。

2.加权轮叫（Weighted Round Robin）（简称wrr)

调度器通过“加权轮叫”调度算法根据真实服务器的不同处理能力来调度访问请求。这样可以保证处理能力强的服务器能处理更多的访问流量。调度器可以自动问询真实服务器的负载情况，并动态地调整其权值。

3.最少链接（Least Connections）(LC)

调度器通过“最少连接”调度算法动态地将网络请求调度到已建立的链接数最少的服务器上。如果集群系统的真实服务器具有相近的系统性能，采用“最小连接”调度算法可以较好地均衡负载。

4.加权最少链接（Weighted Least Connections）(WLC)

在集群系统中的服务器性能差异较大的情况下，调度器采用“加权最少链接”调度算法优化负载均衡性能，具有较高权值的服务器将承受较大比例的活动连接负载。调度器可以自动问询真实服务器的负载情况，并动态地调整其权值。

5.基于局部性的最少链接（Locality-Based Least Connections）(LBLC)

“基于局部性的最少链接”调度算法是针对目标IP地址的负载均衡，目前主要用于Cache集群系统。该算法根据请求的目标IP地址找出该目标IP地址最近使用的服务器，若该服务器是可用的且没有超载，将请求发送到该服务器；若服务器不存在，或者该服务器超载且有服务器处于一半的工作负载，则用“最少链接” 的原则选出一个可用的服务器，将请求发送到该服务器。

6.带复制的基于局部性最少链接（Locality-Based Least Connections with Replication）(LBLCR)

“带复制的基于局部性最少链接”调度算法也是针对目标IP地址的负载均衡，目前主要用于Cache集群系统。它与LBLC算法的不同之处是它要维护从一个目标 IP地址到一组服务器的映射，而LBLC算法维护从一个目标IP地址到一台服务器的映射。该算法根据请求的目标IP地址找出该目标IP地址对应的服务器组，按“最小连接”原则从服务器组中选出一台服务器，若服务器没有超载，将请求发送到该服务器；若服务器超载，则按“最小连接”原则从这个集群中选出一台服务器，将该服务器加入到服务器组中，将请求发送到该服务器。同时，当该服务器组有一段时间没有被修改，将最忙的服务器从服务器组中删除，以降低复制的程度。

7.目标地址散列（Destination Hashing）(DH)

“目标地址散列”调度算法根据请求的目标IP地址，作为散列键（Hash Key）从静态分配的散列表找出对应的服务器，若该服务器是可用的且未超载，将请求发送到该服务器，否则返回空。

8.源地址散列（Source Hashing）(SH)

“源地址散列”调度算法根据请求的源IP地址，作为散列键（Hash Key）从静态分配的散列表找出对应的服务器，若该服务器是可用的且未超载，将请求发送到该服务器，否则返回空。

9. 最短的期望的延迟（Shortest Expected Delay Scheduling SED）(SED)

基于wlc算法。这个必须举例来说了

ABC三台机器分别权重123 ，连接数也分别是123。那么如果使用WLC算法的话一个新请求进入时它可能会分给ABC中的任意一个。使用sed算法后会进行这样一个运算

A(1+1)/1

B(1+2)/2

C(1+3)/3

根据运算结果，把连接交给C 。

10.最少队列调度（Never Queue Scheduling NQ）(NQ)

无需队列。如果有台 realserver的连接数＝0就直接分配过去，不需要在进行sed运算

LVS配置流程框架

拓扑信息收集

Director端配置过程

RealServer端配置过程

测试过程

LVS集群拓扑信息表

角色：Client, Router, VS, RS

拓扑信息：CIP, GWIP, DIP, VIP, RIP,……

IP规划：

Hostname ip

VIP 192.168.1.250

MASTER 192.168.1.202

BACKUP 192.168.1.204

REALSERVER1 192.168.1.201

REALSERVER2 192.168.1.203

配置LVS集群的过程

确认内核支持ipvs

执行命令：modprobe -l | grep ipvs

确定集群架构、类型、角色/成员、拓扑，确定各成员的网络配置参数。

网络环境的配置

专用网络设备：路由器/防火墙/交换机

集群各成员主机配置信息

使用ifconfig或者ip命令配置网络接口

在各个角色主机上跑tcpdump抓包，分析报文通径。这一方法通常用于分析LVS配置中的故障。

前端(FrontEnd)调度器Director的配置

NAT:

外网VIP配置，将作为对外公开的可访问接口

内网DIP配置，将作为后端RealServer的GWIP

内核IP转发参数设置

负载均衡服务、服务器及规则配置

基于ipvsadm脚本的方法

基于keepalived配置文件的方法

DR:

先在对外公开的可访问接口上配置DIP，然后在该接口上创建子接口配置VIP

内核参数设置

负载均衡服务、服务器及规则配置

基于ipvsadm脚本的方法

基于keepalived配置文件的方法

TUN:同DR，但要注意配置重点在隧道设备。