

Linux 集群技术研究

屈 钢¹, 邓健青², 韩云路³

(1. 中国科学院 计算所, 北京 100080; 2. 武汉大学 计算机学院, 湖北 武汉 430072; 3. 吉林大学, 吉林 长春 130000)

摘 要: 面对网络的快速发展, 网络服务器需要具备提供大量并发访问服务的能力。利用服务器集群技术可以有效地解决这个问题。介绍了集群技术的基本原理, 并在 Linux 环境下提出了具体的解决方案和关键技术。

关键词: 集群; 可用性; 负载均衡; Linux

中图法分类号: TP393. 02 文献标识码: A 文章编号: 1001-3695(2005)05-0100-02

Research on Linux Cluster Technology

QU Gang¹, DENG Jian-qing², HAN Yun-lu³

(1. Institute of Computer Technology, Chinese Academy of Science, Beijing 100080, China; 2. School of Computer, Wuhan University, Wuhan Hubei 430072, China; 3. Jilin University, Changchun Jilin 130000, China)

Abstract: Internet server should deal with a lot of concurrent request from users because of the sharply increasing of Internet network. Implement cluster will solve this problem effectually. This paper introduces the principle of implement cluster and puts forward the solution problem and key technology.

Key words: Cluster; Availability; Load Balance; Linux

1 概述

网络的高速发展, 带来了服务器的线形增长, 当访问需求增长时, 服务器会在很短的时间内超载, 集群技术由此应运而生。目前, 越来越多的企业采用 Linux 操作系统, 提供 Web、邮件、文件存储、数据库等服务。随着人们对 Linux 服务器依赖的加深, 对其可靠性、负载能力和数据能力也备加关注。Linux 集群技术以低廉的成本, 实现了这些需求。Linux 可以运行于极为普及的 PC 机上, 不需要购买昂贵的 UNIX 硬件设备。在几台运行 Linux 的 PC 机上, 只要加入相应的集群软件, 就可以组成具有很强可靠性、负载能力和计算能力的 Linux 集群。

1.1 集群的定义

集群就是通过高性能网络将多台同构或异构的计算机连接起来并协同完成特定任务的计算机群, 它们对外部的表现就像一个系统在工作。传统的 UNIX 的集群产品, 如 TruCluster Server 之类, 价格十分昂贵, 而 Linux 集群产品具有优良的性能价格比。

1.2 集群的类别

集群可以分为三类: 高可用性集群、负载均衡集群、超级计算集群。

(1) 高可用性集群。高可用性集群的主要功能就是提供 7 × 24 小时不间断服务。有许多应用程序都必须一天 24 小时不停运转, 如工业控制系统、数据交换机、军事监测系统、电信业务支持系统等。对这些应用程序来说, 暂时的停机和重启都会导致数据的丢失或灾难性的后果。高可用性集群通常是由

多个节点构成, 当单个节点出现故障时, 系统其他节点接替服务, 整个系统仍能继续正常对外提供服务, 其设计目的就是要最大限度地减少服务中断时间。

(2) 负载均衡集群。第二类集群是负载均衡集群, 目的是提供与节点个数成正比的负载能力, 将系统的整体负载合理地分配到各个节点上, 使得每个节点都不会因为超负荷工作而崩溃; 并且系统整体负载需求大于系统整体负载能力时, 可以通过增加系统节点, 平滑地拓展系统负荷能力。这种集群非常适合提供大访问量的 Web 服务。

(3) 超级计算集群。第三类集群是超级计算集群, 它通过将多台机器连接起来同时处理复杂的计算问题。按照计算关联程度的不同, 又可以分为两种: 任务片方式, 要把计算任务分成任务片, 再把任务片分配给各节点, 在各节点上分别计算后再把结果汇总, 生成最终计算结果。 并行计算方式, 节点之间在计算过程中大量地交换数据, 可以进行具有强耦合关系的计算。这两种超级计算集群分别适用于不同类型的数据处理工作。有了超级计算集群软件, 企业利用多台 PC 机的组合就可以完成通常只有大型服务器才能完成的计算任务。

2 Linux 集群方案

根据不同的应用需求, Linux 集群有三种类别的集群方案: 高可用性集群方案、负载均衡集群方案、超级计算集群方案。

2.1 Linux 高可用性方案

(1) 高可用性集群系统架构。高可用性集群系统中不同的服务器承担不同的任务, 当其中一个服务器发生故障时, 系统根据设定的条件, 将发生故障的服务器的任务转移给另外一台服务器, 对于最终用户来说, 并没有反映出系统故障, 系统的

可用性得到提高。

(2) Linux 高可用性集群软件。Linux 平台的集群软件有 LifeKeeper for Linux , Convolvo, TurboHA 以及 Motorola 的 HA-Linux 等。

以 TurboHA 为例, 其软件功能如下:
系统配置。在 Dw. cfg 系统文件中设置系统的配置信息。

通信监测。在主备服务器之间设置一个通信交互区交流系统运行状态。TurboHA Server 根据服务器写在这块区域上的心跳信息来判断服务器是否正常运行。当 TurboHA Server 检测到工作主机心跳停止时, 会通过其他的监控路径确认这台服务器是否真的出现异常。如果确定, 则由另一台服务器接管这台服务器的工作。

虚拟 IP。整个 HA 系统对外提供一个虚拟的 IP 地址, 系统的客户端通过虚拟地址实现与主机之间的通信, 无论系统是否发生切换, 虚拟地址始终指向工作主机, 从而达到了保障网络主机服务不间断的目的。

2.2 负载均衡集群方案

(1) 负载均衡集群系统架构。负载均衡的解决方案中, 若干台服务器做同样的工作, 这样一来以前由一台服务器来做的工作已经分配给多个服务器来做, 整个系统的处理能力得以提高。在集群中有一个主控节点, 称为高级流量管理器(ATM)。用户请求发送到 ATM 上, 在 ATM 上绑定系统服务对外的 IP 地址。ATM 把接收到的请求平均发送到各服务节点上, 服务节点接收到请求之后, 直接把相应的处理结果发送给用户。这样一来, 假如在 1s 内有 1 000 个请求, 而集群中有五个服务节点, 则每个节点将处理 200 个请求。这样, 对于外界来说, 相当于有一台五倍速度的高速计算机在处理用户的访问。这也就是真正意义上的负载均衡。

(2) 负载均衡集群软件。Linux 平台下的集群软件有 Turbolinux Cluster Server, Linux Virtual Server 等。

以 Turbolinux Cluster Server 为例, 其软件功能如下:
配置管理。高级流量管理器(ATM) 读入配置文件, 进行流量路由的初始化。
流量管理。主 ATM 接收所有从客户端来的连接请求, 并利用设定的流量转发机制和均衡负责方案将这些请求转发到合适的集群节点。

虚拟 IP。整个集群系统提供一个统一对外的单一虚拟 IP 地址, 由一个虚拟 IP 地址来对外代表整个集群。

服务器检测。检测服务节点的性能、负载情况、服务进程、服务运行环境。

2.3 超级计算集群方案

(1) 超级计算集群系统架构。超级计算集群是并行计算的基础, 以解决复杂的科学问题。它可以使集群系统通过高速连接来链接的一组单处理器或双处理器微机, 并且在公共消息传递层上进行通信以运行并行应用程序。因此, 所谓的廉价的 Linux 超级计算机, 实际是一个 Linux 计算机群集, 其处理能力与真正的超级计算机相当, 但 Linux 集群系统的价格比超级计算机便宜很多。

(2) 超级计算集群软件。支持 Linux 集群的软件及系统有 EnFuzion, Beowulf 等。

Beowulf 软件的主要功能包括: 分发任务到各个计算节点; 监测任务运行的情况; 监测故障及故障恢复; 控制任务队列等。

3 Linux 集群关键技术

实现 Linux 集群所需的关键技术有 IP 虚拟服务技术、负载均衡调动算法、节点间通信接口体系、系统监测技术等。

3.1 IP 虚拟服务技术

所谓 IP 虚拟服务技术, 是指整个集群系统提供一个统一对外的单一虚拟 IP 地址, 由一个虚拟 IP 地址来对外代表整个集群, 使用户感觉到是由一台大型计算机在向其提供服务。

IP 虚拟服务技术依赖于 IP 伪装机制。一般 IP 伪装机制由 NAT(Network Address Translation) 技术来实现。NAT 支持静态网络地址翻译和动态网络地址翻译。NAT 能将 IP 地址从一组映射到另外一组, 如果这种映射关系是 M-M 的, 则称之为静态网络地址翻译; 如果映射是 M-N(M > N) 的, 则叫做动态网络地址翻译。IP 伪装机制就是一种 M-1 的动态网络地址翻译, 它能够将多个内部网中的 IP 地址映射到一个与 Internet 相连接的外部网 IP 地址上, 使整个网络虚拟为一台大型计算机, 用一个 IP 地址向外界提供服务。

通过 NAT, 与外网连接的调度服务器重写请求报文的目标地址, 根据负载均衡调度算法, 将请求分派给后端的服务器; 后端服务器的响应报文通过调度服务器时, 报文的源地址被重写, 再返回给客户, 完成整个请求 - 响应过程。这种通过 NAT 将一组小型计算机构成一个高性能的大型虚拟服务器的技术, 称之为 VS/NAT 技术(Virtual Server via Network Address Translation)。

3.2 负载均衡调度算法

负载均衡集群的调度服务器通过调度模块, 将用户的请求转发到某一个后端服务器上。负载均衡调度算法包括轮叫算法、连接算法、系统监测算法等。

(1) 轮叫算法: 是指让调度服务器将外部请求按顺序轮流分配到集群中的后端服务器上, 而不论服务器上实际的连接数和系统负载。

(2) 连接算法: 是指让调度服务器动态地将网络请求调度到已建立的连接数最少的服务器上。

(3) 系统监测算法: 是指通过监测后端服务器的综合性能指标, 如 CPU 使用率、空闲内存容量等, 加权评估系统总体的性能, 让调度服务器将外部请求发送到负载最轻的服务器上。

3.3 节点间通信接口技术

在集群节点间可采用的通信接口技术包括: TCP/IP、UDP/IP、磁盘通信区域、VIA(Virtual Interface Architecture) 等。

(1) TCP/IP 及 UDP/IP。它们是应用最广泛的通信方式。在高可用性集群中, 当系统监测进程发现磁盘通信区域的心跳信号消失后, 会使用 TCP/IP 或 UDP/IP 协议通过网络确认被监测服务器的状态。在负载均衡集群中, 可使用 TCP/IP 承载 HTTP, FTP 信息, 实现前端服务器和后端 Web 服务器之间的通信; 使用 UDP/IP 承载视频、音频信息, 实现前端服务器和后端媒体服务器之间的通信。

(2) 磁盘通信区域。这种方式应用于高可用性集群, 在主备服务器之间的磁盘阵列上划分一块公共通信(下转第 107 页)

志, 即使攻击者获得原有的一份签名 $(\{ C^{(i)} \}_{i=1}^l, T)$, 他可以将 T 变成现有的时间, 但他无法改掉 M 中的 $h(g^k \bmod n, T)$, 因此攻击不会成功。

4.2 效率分析

(1) 本方案在运用多素数 RSA 算法进行解密时需要进行 s 次(s 为私钥的个数) 指数长度为 n/s bit 位的指数模运算 $C^d \bmod n$ 。传统 RSA 在用基本 BR 算法计算时所用时间复杂度为 $O(\log d \log^2 n)^{[1]}$, 当 d 很大与 n 同数量级时, 复杂度可以认为是 $O(\log^3 n)$ 。这样, 多素数 RSA 改善传统 RSA 计算效率是显而易见的。在效率上提高的倍数为: $\alpha = \frac{(\log n)^3}{s(\log(n/s))^3} = s^2$ 。例如当密钥个数 $s = 2$ 时, 效率大约是传统计算方法的四倍。文献[5] 给出实际实验结果: 当 $n = 768$ bits 时 $\alpha = 3.24$, $n = 1\,024$ bits 时 $\alpha = 3.32$, 当 $n = 2\,048$ bits 时, α 可以提高到 3.47 倍^[5]。而我们的方案比文献[5] 的效率平均提高了 1.29 倍左右。

(2) 本方案在运用中国剩余定理中计算逆元时, 巧妙地通过一个多项式运算来代替其中一个逆元的求解, 进一步提高了运算效率。即把原来 s 个密钥需要计算 s 次逆元, 减少为 $s - 1$ 次。特别是当 s 不是很大时, 效果是显著的。例如当 $s = 3$ 时, 大约节省了 33% 的计算量。

效率分析表明, 本方案采用多个短密钥来代替传统的单密钥, 并且在运用中国剩余定理时巧妙地求解一个逆元, 减少了计算量, 提高了效率, 非常适合基于智能卡这样的资源受限的加密系统。

5 结束语

本文提出了一种新的多密钥 RSA 的认证加密方案, 加密的同时考虑到了消息认证, 具有文献[5] 的高效率的优点, 并且巧妙地运用中国剩余定理来减少求逆元的次数, 使效率得到

进一步的提高。同时对恢复消息进行认证, 以及加入时间标志以防止重发攻击, 使通信更加安全, 非常适合在智能卡上的加密认证领域的应用。

参考文献:

[1] L Rivest, A Shamir, L Adleman. A Method for Obtaining Digital Signatures and Public Key Cryptosystems[J]. Communications of the ACM, 1978, 21(2): 120-126.

[2] Arjen K Lenstra, Eric R Verheul. Selecting Cryptographic Key Sizes [J]. Journal of Cryptology, 2001, 14(4): 255-293.

[3] D M Gordon. A Survey of Fast Exponentiation Methods[J]. Journal of Algorithms, 1998, 27: 129-146.

[4] D Boneh, H Shacham. Fast Variants of RSA[R]. RSA Laboratories Cryptobytes, 2002, 5(1): 1-8.

[5] Cesar A M Paixao. An Efficient Variant of the RSA Cryptosystem [EB/OL]. Institute of Mathematics and Statistics University of Sao Paulo, Brazil, <http://www.ime.usp.br/~capaixao/paper.pdf>, 2003.

[6] M J Hinek, M K Low, E Teske. On Some Attacks on Multy-prime RSA [EB/OL]. The Centre for Applied Cryptographic Research (CACR) at the University of Waterloo, Department of Combinatorics and Optimization, Ontario, Canada. <http://www.cacr.math.uwaterloo.ca/techreports/2002/corr2002-11.ps>, 2002.

[7] M Joye, P Paillier. How to Use RSA; or How to Improve the Efficiency of RSA Without Loosing Its Security[C]. Paris, France: 2002 Information Security Solutions Europe Conference (ISSE2002), 2002.

[8] D Boneh, X Ding, G Tsudik, *et al.* A Method for Fast Revocation of Public Key Certificates and Security Capabilities[C]. Washington, D. C., USA: the 10th USENIX Security Symposium, 2001. 297-308.

[9] X Ding, G Tsudik. Simple Identity-based Encryption with Mediated RSA[C]. San Francisco, CA, USA: 2003 Cryptographer's Track RSA Conference, 2003. 193-210.

作者简介:

章磊(1977-), 男, 浙江衢州人, 硕士, 研究方向为计算机网络与安全; 卢建朱(1965-), 男, 湖南郴州人, 副教授, 博士, 研究方向为多媒体中的数据处理和通信技术、计算机网络与安全; 凌捷(1964-), 男, 广东梅州人, 教授, 博士, 研究方向为信息安全技术、视频图像处理; 李家兰(1977-), 男, 广东茂名, 人, 硕士。

(上接第 101 页) 区域, 主服务器上的被监测进程向此区域写入心跳信号, 备服务器上的监测进程读这块磁盘区域上的心跳信号, 以此方式来判断主服务器的运行状态是否发生异常。

(3) VIA。在超级计算集群中, 对各个计算节点的通信效率有非常高的要求。基于传统的 TCP, UDP 网络协议的 Socket 原语通信方式, 带来了大量的集群系统开销, 不能很好地满足科学计算集群高效率、低延迟的通信性能要求。在此背景下, 业界开始了轻量级通信协议的研究, 并取得了很大进展, 虚拟接口体系结构(VIA) 在此基础上产生了。VIA 是一种轻量级的网络接口体系结构规范, 它比 TCP/IP 协议的系统开销要小很多。TCP/IP 协议的实现需要较多的操作系统干预, 由此引入大量的系统开销, 如系统调用、数据拷贝、复杂的协议栈处理、关键路径的大量干预等。而 VIA 避免了操作系统的大量干预以减轻由此引起的系统开销。在 VIA 体系结构中, 操作系统的干预只发生在通信信道的分配和连接建立的过程中, 在随后的消息传递过程中, 网络接口直接从发送进程的用户空间获取数据和目标地址, 并直接传送到接收进程的用户空间缓冲区。VIA 避免了操作系统在数据收、发过程中的拷贝开销, 同时也减轻了复杂协议栈处理引起的系统开销。

4 总结

Linux 集群技术在各个领域有着广泛的应用。其行业及应用目标如表 1 所示。

表 1 Linux 集群技术应用行业及目标

	电信领域	金融领域	游戏领域	CAD 设计	气象预测	科学研究
高可用性集群	实时计费系统	交易处理	图形对象数据库	工程数据库服务	气象信息数据库	科技信息数据库
负载均衡集群	网上营业系统	网上银行	网上游戏服务器	图形对象发布	气象信息检索服务	科技信息检索服务
超级计算集群	用户信用度分析	风险分析	动画着色	部件模型	气象概率分析	DNA 模型

参考文献:

[1] 沈中林, 金宝鼎, 扬明. 利用集群技术构建 Linux Virtual Server [J]. 计算机应用研究, 2000, 17(3): 30-32, 68.

[2] 姚耀文. 基于 Linux 的服务器群集方案[J]. 计算机工程, 2001, 27(4): 139-141.

作者简介:

屈钢, 信息产业部认证项目经理, 在职研究生, 有丰富的项目开发经验, 已经在国内、国际会议及核心期刊上发表多篇论文; 邓健青(1971-), 女, 硕士, 研究方向为计算机网络及应用; 韩云路, 铁通通信信息有限责任公司吉林分公司副总经理, 高级工程师, 博士研究生。