文章编号:1008-0570(2008)12-3-0109-03

一种分布式集群系统

A wide-area distributed cluster

(厦门大学) 李建敏 李翠华
LI Jian-min LI Cui-hua

摘要: 本文研究了广域网分布式集群的设计。与 LVS 相比,这种集群的结构虽然复杂,但是能够克服传统集群的局限性,提供质量好,容量大,性价比高的解决方案。根据一种称之为 CG 树的分布式集群的抽象模型,作者设计了自己 CG 树算法。这种方案能有效的减少网络节点之间的通信开销,并能在集群系统出现故障时自动进行调整,保证了高可用性。模拟实验结果表明,该系统具有很好的性能。

关键词: 分布式; 集群; CG 树; LVS
中图分类号: TP302.1 **文献标识码:** A

Abstract: The paper studies the designs of wide-area distributed cluster. Although these kinds of cluster have more complex structure, they can overcome limitations of traditional cluster and provide resolutions with good quality, large capacity and high performance compared to LVS. Base on a model of distributed cluster, named CG tree, the author design his algorithms to manipulate CG trees. The scheme can efficiently reduce the cost of intercommunication between different networks, and can automatically adjust tree to ensure high availability when cluster faults occur. The result of simulated experiments indicates that the system has quite good performance.

Key words: Distributed cluster; CG tree; LVS

技术创新

1 引言

随着社会的发展,Internet 技术的进步, 网民数量在迅速的增加,网络视频服务也越来越流行。网络媒体服务商也不断的出现。但视频的信息量很大,他们提供的服务往往只能满足少量的用户同时使用,当在线用户超过一定数量时,服务的质量就会很差。

这个问题归结起来,主要是两个原因。一,服务器性能的限制;二,网络带宽的限制。解决的办法很多,但这两方面的问题往往分开来解决。针对服务器性能差,可以采用高性能服务器或者架构集群的办法,针对网络带宽的限制,采用专用的网络接口卡,提高网络速度,这种方法效果不是很好。

服务商采用的集群基本上都是集中在一个网络内,用专有的大容量带宽对外提供服务,例如 Real 公司就使用由 20 台服务器组成的 LVS 集群,为其全球用户提供音频视频服务。但很多小的服务商就没有能力采用如此昂贵的办法。即使是大服务商,在当用户规模不断扩大时,也面临网络的问题。

采用集群做服务器的方法, 使服务器的性能问题得到解决,但是服务器集中在一个网络内,就使得服务器的网络带宽成为系统对外服务的瓶颈。许多服务商解决这个问题的办法是 :将媒体资源做成镜像,放置于不同的网络,同时给用户提供多个链接,例如 :PPLIVE,PPSTREAM 等播放软件就采用这样的办法。这就造成很大的冗余,而且也给用户造成不便,用户往往要多次尝试才能知道哪个链接的效果好。

采用 LVS 的办法架构的多媒体服务器, 一般也是集中式的。LVS 的 loadballance 要做流量分配的任务,负担很重,当集群的规模扩大时,loadballance 就有可能成为系统的瓶颈。如果架构于不同的网络,问题将会更加突出。因此 LVS 的扩展限制也是很大的,也没能解决网络带宽的问题。

2 系统的提出

为解决以上的问题,下面我们举个典型的城市宽带小区的例子做说明。该城市由三个区组成,每个区各自包含若干个节点,它们之间通过高速网络相连。若要为此类宽带小区的用户提供媒体视频服务,可以考虑以下几种方案:

(1) 集中式方案,媒体服务端使用高性能专用服务器,采用高速网络接入,这种结构的优点是管理方便,但显然有上文提到的两大问题。

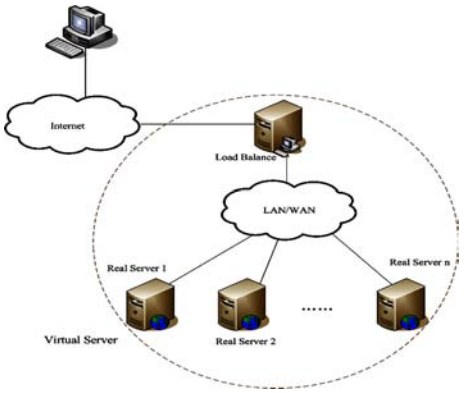

图 1 典型的 LVS 集群结构

(2) 采用 LVS 的解决方案,该方案的特点是构建简洁,性能

李建敏: 硕士研究生

基金项目: 国家重点基础研究发展计划 (973 计划) 项目(No. 2007CB311005);国家 863 计划项目(2006AA01Z129);福建省自然科学基金计划资助项目(A0710020);厦门大学 985 二期信息创新平台项目

价格比好,能基本满足中小规模的视频服务应用,且具有一定的高可用性,但该集群方案对这类结构来说,其服务器仍然是集中式的,而且服务器池采用单层线性结构,可扩展性有限,采用这种结构的服务器节点通常架设在内部局域网中。同时 LVS 这种结构对服务端的网络性能要求较高,如果在更大范围内应用,如对跨网络的分散节点进行组织,则很难有理想的效果。

(3) 采用纯粹分布式方案,如在每个宽带小区节点所在的位置放置一个媒体服务器节点,该节点可以使用普通主机,每个节点能独立提供媒体服务,并能够根据实际网络环境对附近的用户提供服务,将这些分散的节点组合起来可以形成一个功能强大的媒体点播环境,但它的缺点是为了做出合理决策,媒体服务器节点之间需要两两通信,特别是跨网络节点之间,其通信开销将是巨大的。

一种比较理想的方案是在上述分布的基础上采用分级分层的媒体服务器池结构。节点之间不必两两通信,特别是在跨地域的应用系统中,不同网络间的通信开销将大大减小。同时,目前服务提供商往往在不同省份和城市规划不同的系统,造成大量资源浪费,如果能统一规划此类分布式系统,将节省不少成本。

我们认为针对媒体集群采用分级的分布式集群系统构建是一种比较好的方案,因此我们打算设计一种分级的分布式集群管理系统,它能达到高可伸缩性、高可用性、高性价比,同时又能解决不同网络宽带带来的问题。

3 系统概念模型

3.1 CG 树

针对上节提出的分层分级的服务器池的模型,我们给出了这一模型的形式化描述。将这种结构称为 CG (Cluster Group Tree)]树。我们引用了其中的相关定义如下:

组关系

设非空集合 N,定义 N 上的关系 R 具有如下特点:

(1) $\forall b \in N$, $\exists a \in N$,使得 $<a,b> \in Q$

(2) 若 $<a,b> \in R$,则 $<a,a> \in R$

(3) 若 $<a,b> \in R$ 且, $<b,c> \in R$ 则 $a = c$

则称关系 R 为组关系,关系元素<a,b>中 a 称为组长。

同组关系

设非空集合 N 和 N 上的组关系 R, 定义其上的关系 Q 具有如下特点:

(1) $\forall a,b \in N$,若 $<a,b> \in R$ 使得 $<a,b> \in Q$

(2) $\forall a,b \in N$,若 $<a,b> \in Q$ 使得 $<b,a> \in Q$

(3) $\forall a,b,c \in N$,若 $<a,b> \in Q$ 且 $<a,c> \in Q$ 使得 $<b,c> \in Q$

则称关系 Q 为在集合 N 的组关系 R 上的同组关系。

根据同组关系的定义,容易看出,以上的同组关系 Q 是集合 N 上的等价关系。同时,由于集合 N 上的等价关系可以唯一确定 N 的划分, 因为可以说集合 N 上的同组关系 Q 的等价类构成集合 N 的一个划分。

CG 森林

给定结合 N 和关系序列 $R_1,R_2,R_3\cdots,R_m$,其中 R_1 是定义在 N 上的组关系,$R_k+1(1 \leq k \leq m-1)$是定义在 R_k 的所有组长集合上的组关系,记做 $R=<N,R_1,R_2,\cdots,R_m>$,并称 R 为 CG 森林,R_i 的下标 i 成为 CG 森林的层(level).若 R_l 中只有一个元素。则称这样的 CG 森林为 CG 树。

3.2CG 树的意义

有了 CG 树的概念后,一个广义的集群系统就可以很方便的用一颗 CG 树来描述,对集群系统的组织和管理就可以相应的转变为对 CG 树的操作。在典型的集群系统中,后端可以看成是一组服务器节点池,其中在 CG 树中叶子节点代表各个真实的主机节点,它们在逻辑上被划分为多个特定的组 ;非叶子节点与它的其中一个子节点是同一个真实的主机节点,可以说 CG 树中的任何一个非叶子节点都是从它的子节点中选出来的 ;同一个节点可能被包含在多个组中,CG 树中的非叶子节点即为其子节点所在组的组长。

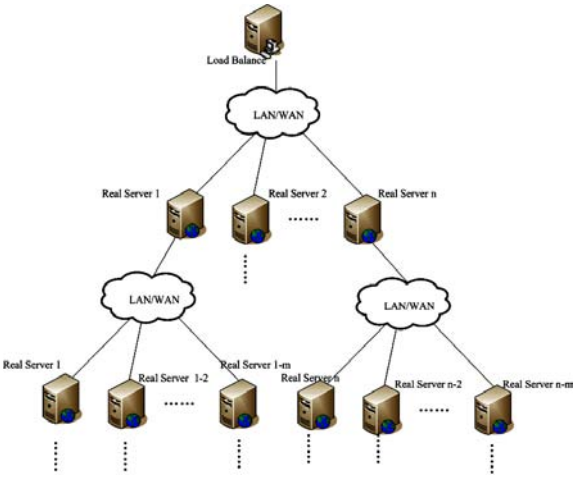

图 2 基于 CG 树的集群结构

3.3 CG 树的实用背景

在实际应用中,CG 树可以应用于跨网络的集群环境。一般来说,距离越远的节点之间通信延迟越大,我们的原则是尽可能将临近的主机(如果同一个局域网)放置于更低层的组,然后按地区和地域分布再逐层划分。如果局域网的主机较多,可以在同一个局域网中再分为多个组,一般不将不同网络的节点放置在底层的组中,这样,节点间在进行跨网络通信时,只通过上层的组长节点进行转发,而不采用两两通信,整个集群系统的通信量将会显著减小。

4 系统设计

4.1 系统结构

根据以上 CG 树的定义,我们可以得到图 2 所示的结构图,也就是本文所提的分层分级的集群结构。

与 LVS 一样,loadballance 是整个集群对外的服务接口。

4.2 通讯规则

心跳(Heart Beat)

集群中的节点定期的通过网络交换信息数据包,这些数据包称为心跳。心跳的典型功能是向对方定期发送一个数据包表明还"活着",实际上,心跳也是一种"通信"机制,只不过它交互的是底层节点提供服务的时候"存活"的消息而已。在典型高可用性集群系统中,心跳消息占据了整个集群系统通信量的绝大部分。在 LVS 中,所有的节点都必须和 loadballance 之间发送心跳,这样导致 loadballance 需要花费大量的时间处理这些心跳。如果底层的节点扩展太多的话, 将会导致 loadballance 疲于处理这些消息而成为整个系统的瓶颈。

在本文的新的分层分级的服务器池结构中,心跳只存在组员节点与组长节点之间, 这样大大减少的了 loadballance 的负担,也减少了心跳在网内的传递。

技术创新

4.3 负载均衡

所有组长节点与 LVS 的 loadballance 相似，都有负载均衡的能力。每个组都相当于一个独立的 LVS。当组长接收到客户对资源的请求时，依据负载均衡算法，这里我们可以采用 LVS 中的任务分配算法，如轮叫调度[5]等方法，将请求分配给某一个组员节点，当然也有可能分配给自己。如果分配到任务的节点仍是一个组长的话，可以继续根据把任务分配给组员。

这样一来的话，就把任务分配的任务从 loadballance 解放出来，分发给各个组长，大大降低了 loadballance 的负担，保证了其稳定性。

4.4 通讯故障处理

当组员节点与组长节点出现通讯故障时：

(1) 对于组长来说，组长在一定时间内没有接收到这个组员的心跳的话，会主动将其从当前组中除名。

(2) 对于这个组员来说，它虽然无法跟组长节点通讯，但是它会主动去测试和同组节点的通信情况，如果和某个节点 A 通信良好的话，这个节点会把自己以及自己的组加入到，以 A 节点为组长的组中。否则的话就不断的给原有组长发心跳，直到有收到组长回复之后，重新加入集群中，提供服务。

当然如果集群服务过程中产生了通讯故障的话，根据以上规则自适应调整之后，集群结构会跟原有结构略不相同，不过这并不影响集群的性能。

5 结论

本文提出了一种新的基于 CG 树的分层分级的集群结构，相比 LVS 的线性结构，虽然结构复杂了些，但是仍有以下优势：①减轻了 loadballance 的负担；②增强了集群的可扩展性；③故障时，集群能够自适应调整，保证了集群的高可用性；

本文作者创新点：(1)提出了一种基于 CG 树结构的集群系统；(2)提出故障时，集群结构自适应调整的策略。

参考文献

[1] 章文嵩.Linux 服务器集群系统 [html]，http://www.linuxvirtualserver.org/zh/.2002-3.

[2]刘维峰.分布式媒体集群的设计与实现 [D].厦门：厦门大学，2005.

[3]张燕，张国平.IPTV 中基于集群技术的 LVS 研究与应用 [J].微计算机信息，2008，5-3：2-4。

[4] Alex Vrenios 著.马朝晖等译.Linux 集群体系结构[M].北京：机械工业出版社，2003

[5] Rajkumar Buyya，高性能集群计算：结构与系统（第 1 卷）[M]，北京：电子工业出版社，2001.

作者简介：李建敏（1984-），男，福建福安人，硕士研究生，主要研究方向数字图像处理。李翠华，博导，教授，主要从事计算机视觉(CV)、视频与图像处理（VIP）、人工神经网络(NNs)、数字水印处理（DWP）等方面的研究工作。

Biography: LI Jian-min (1984 -), Male (Han Nationality), Master, was born in Fujian Province, studies in Xiamen University, researches on digital image process.

(361005 福建厦门 厦门大学信息科学与技术学院) 李建敏

(361005, Coll. of Information Sci. and Techn., Xiamen Univ., Xiamen, Fujian, China) LI Jian-min

通讯地址：(361005 福建厦门 厦门大学信息科学与技术学院) 李建敏

(收稿日期:2008.11.15)(修稿日期:2008.12.13)

(上接第 41 页)

[2]袁红林，徐晨，章国安.TOSSIM:无线传感器网络仿真环境[J].微计算机信息，2006,7-1:154-15

[3]Malan, D., Welsh, M., Smith, M.: A Public-Key Infrastructure for Key Distribution in TinyOS Based on Elliptic Curve Cryptography, IEEE SECON (2004)

[4]Du, W., Deng, J., Han, Y.S., Chen, S., Varshney, P.: A Key Management Scheme for Wireless Sensor Networks Using Deployment Knowledge. IEEE INFOCOM (2004)

[5]Miller, V.S.: Use of Elliptic Curves in Cryptography. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 417 - 426. Springer, Heidelberg (1986)

[6]Carman, D.W., Kruus, P.S., Matt, B.J.: Constraints and approaches for distributed sensor network security. NAI Labs Technical Report 00-010 (September 2000)

[7]Chan, A.C., Rogers, Sr., E.S.: Distributed Symmetric Key Management for Mobile Ad hoc Networks IEEE INFOCOM (2004)

[8]Eschenauer, L., Gligor, V.D.: A key-management scheme for distributed sensor networks.In: Proceedings of the 9th ACM conferenceon Computer and communications security,Washington, DC, USA (November 18 - 22, 2002)

[9]Sun, Y., Liu, K.J.R.: Scalable Hierarchical Access Control in Secure Group
Communications, IEEE INFOCOM (2004)

[10]Steiner, M., Tsudik, G., Waidner, M.: Key Agreement in Dynamic Peer Groups. IEEE Transactions on Parallel and Distributed Systems 11(8), 769 - 780 (2000)

作者简介：胡松(1983-)：汉族，湖北人，硕士生．研究方向：无线传感器网络．樊晓平(1961-)：汉族，浙江人，博士，教授、博士生导师．研究方向：无线传感器网络、智能控制、机器人控制．刘少强(1964-)：汉族，湖南人，博士，副教授，硕士生导师．研究方向：无线传感器网络

Biography: HU Song (birth year-1983),male(han ethnic),Hubei, School of Information Science and Engineering ,CSU, MS Candidate, Research area:WSN

(410075 长沙 中南大学信息科学与工程学院) 胡 松 樊晓平 刘少强

(School of Information Science and Engineering,CSU, Changsha 410075) HU Song FAN Xiao-ping LIU Shao-qiang

通讯地址：(410075 湖南省长沙市中南大学铁道校区电子楼 419)胡 松

(收稿日期:2008.11.15)(修稿日期:2008.12.13)

技术创新