分组技术及其在集群中的应用

Application of Teaming Technique in Cluster

姜　游　（中国石油化工股份有限公司 南京石油物探研究所 江苏 南京 210014）

摘　要：在集群系统中，I/O 节点担负着数据存储和分发的角色。针对 I/O 密集型应用，I/O 节点网络带宽常常无法满足需要而成为整个系统的瓶颈。分组技术可以方便、高效地增加 I/O 节点带宽，并提供冗余保护。本文简要介绍了分组技术原理、模式，并通过在集群 I/O 节点上安装 intel 公司提供的 iANS 具体实现。最后给出了性能评价。此技术在实际应用中已取得良好效果。

关键词：分组 teaming 集群 iANS 带宽 性能测试

由于集群（Cluster）**在性能价格比、可扩展性、可用性等方面的巨大优势，使其广泛应用于国防、气象、石油勘探、医学、生物工程、天体物理、工业设计等各个领域。对于采用** Server/Client **模式通过** NFS **实现磁盘空间共享的集群架构而言，网络可用性**（Availability）**和网络带宽容量**（Network Bandwidth Capacity）**就显得十分重要。以石油勘探数据处理为例，一个工区处理的原始数据量以** TB **为单位，加上处理过程产生的** 4 **倍于原始数据的中间数据，其规模十分庞大。普遍而言，**S/C **模式的基本架构如图一所示，数据一般都由** I/O **节点负责存储和分发，其结果通常是** I/O **节点在扮演着瓶颈的角色，因此如何提高** I/O **节点的网络带宽性能就成了急需解决的问题。其中既方便又廉价的解决方式便是将多块网络适配器分组**（Teaming）**，以提高网络可用性、增加网络带宽容量以及实现负载均衡**（Adaptive Load Balancing，ALB）**、适配器冗余容错**（Adapter Fault Tolerance，AFT）**等功能。**

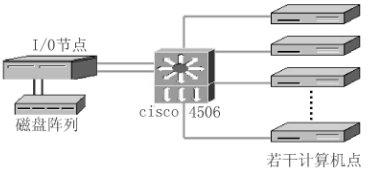

图1　S/C 模式架构

1　分组（Teaming）技术介绍

1.1 目前流行的分组技术

网络适配器分组就是将在一台服务器上的多个物理适配器通过软件绑定成为一块虚拟的适配器，让这些适配器并行链接聚合为一个逻辑链路工作，对于外部网络而言这就是一个单独的以太网接口设备。目前与绑定相关的技术有 SUN **公司的链路聚集**（Trunking）**技术，**Cisco **公司的** Fast Ether Channel（FEC）**技术，**Intel **公司的** Advanced Network Service（iANS），**以及** Alteon **的** Fault Tolerance。**在** Windows **平台上，**Intel **公司提供了** PROSet **工具来实现，**Broadcom **公司提供了** BASP（Broadcom Advanced Server Program）**来实现。在** Linux **的** 2.4 **内核中已经加入了绑定技术，并应用在** Beowulf **集群上。**

有必要简单叙述一下分组的原理。通常情形下，网卡只接收硬件地址（MAC Address）**是自身** MAC **地址的包，其他的包都被丢弃。当处于混杂**（Promisc）**模式时，网卡将接收网络上所有的包。分组技术使多块网卡均处于混杂模式，使其接收特定** MAC **目的地址的包，然后交由分组驱动程序处理，以达到网卡绑定的目的。**

1.2 分组模式

iANS（intel Advanced Networking Services）**是** Intel **公司提供的在** Linux **平台下实现多块网卡分组的软件包，目前最新版本为** 3.4.3，**以它为例简要介绍一下目前流行的分组技术支持的主要分组模式：**

1.2.1 适配器冗余容错(Adapter Fault Tolerance,AFT)

这是 iANS 创建分组时的默认模式。在该分组模式下,每次都只有一块适配器被激活。在主连接因为适配器、双绞线或端口等各种原因发生故障时,组中的备用适配器便自动激活。如果该组中没有指定首选适配器,则在故障恢复后,控制权仍然留给当前被激活的适配器。如果该组中指定了首选适配器,则在故障恢复后,控制权自动交还给首选适配器。

1.2.2 交换机冗余容错(Switch Fault Tolerance,SFT)

该模式支持两块网络适配器分别连接到不同的交换机。但是需要交换机打开生成树协议(STP)来防止循环。每次均只有一块适配器被激活。

1.2.3 自适应负载均衡(Adaptive Load Balancing,ALB)

该模式支持 2 – 8 块适配器形成的组,支持组成员之间的混合速度/双工模式设置,且对交换机无特殊配置要求。在该分组模式下,所有适配器均被激活,在单点由于各种原因发生故障时,组中的其余适配器继续工作,故能起到失效保护作用。该模式包括了接收负载均衡(Receive Load Balancing ,RLB),默认情况下 RLB 将被启用,此时所有被激活的适配器将接收 IPv4 通信包。

1.2.4 静态链路聚合(Static Link Aggregation,SLA)

该模式支持 2 – 8 块适配器形成的组以增强传输和接收的带宽容量,并包括 AFT 和 ALB,但需要交换机支持 Intel Link Aggregation 或 Cisco 的 FEC/GEC,且交换机的生成树协议必须关闭。

1.2.5 IEEE 802.3ad

该模式适用于融合到 Cisco FEC 方式中的 IEEE 标准。iANS 对动态 802.3ad 分组的支持和 SLA 分组类似。

通过上述 5 种分组模式的分析,笔者认为能扩大带宽容量、实现负载均衡、起到失效保护的 ALB 模式最符合我们的需求。

2 分组技术实现过程

以 iANS 为例,简要说明一下分组技术的实现过程。

iANS 软件要求 Linux 内核版本为 2.4.7(或更高),同时要求分组中至少有一块 Intel 的 PRO/100 或 PRO/1000 网络适配器,且 PRO/100 驱动程序版本为 2.0.x(或更高),PRO/1000 驱动程序版本为 4.2.x(或更高)。该软件包同时支持基于安腾的系统。需要注意的是,对于 MVT(Multi – Vendor Teaming,多供应商分组)的支持,除了绑定的网卡中至少有一块 Intel 的网络适配器外,还要求非 Intel 的网络适配器能正常驱动。

2.1 iANS 的安装

以 iANS – 3.4.3a.tar.gz 包为例,安装过程如下:

(1)确认网络适配器工作正常,并且驱动程序版本符合要求。

(2)获得 iANS。最新的 iANS 可以通过 Intel 的支持网站 http://support.intel.com 获得。

(3)以 root 用户身份登陆。

(4)解开 iANS – 3.4.3a.tar.gz:

[root@io2 root]# tar xzvf iANS – 3.4.3a.tar.gz

解开后生成一个名为 iANS – 3.4.3a 的目录。

(5)进入 iANS – 3.4.3a/src 目录进行编译:

[root@io2 root]#make

(6)进行安装:

[root@io2 root]#make install

这样就结束了整个安装过程。配置 iANS

2.2 配置 iANS

配置 iANS 有如下 3 种方式:

a. PROCfg 实用程序,这是一种简单的配置工具,可以减少命令行输入量。

b. 使用 ianscfg 工具手工修改,建议高级用户使用。

c. 使用 ianstool 脚本,建议新用户使用。

下面以 ianstool 脚本方式为例进行配置。

(1)确保网卡已经加载所需的驱动,并停用需要进行绑定的网络适配器。在笔者的实例中停掉了 eth1 和 eth2:

[root@io2 root]# ifconfig eth1 down

[root@io2 root]# ifconfig eth2 down

(2)加载 iANS 模块:

[root@io2 root]# insmod ians

(3)运行 ianstool 配置 iANS:

[root@io2 root]# ianstool

①Add team
②Del team
③View team
④Configure team
⑤Save topology
⑥Exit
Select menu option: 1
//选择 1 进行添加组操作
Enter team name: ioteam
//为新添加的组命名,本例中命名为 ioteam
Select teaming mode:
One of [AFT SFT ALB SLA 802.3ad NONE] (default – AFT): ALB
Do you want RLB to be enabled? ([Y]/n): Y
//选择分组模式,本实例中,笔者选择了 ALB,并激活了 RLB
Select VLAN mode:
One of [none static GVRP] (default – none):
//我们不使用 VLAN,直接回车即可
Enter member name, <Enter> to end:
One of [eth0 eth1 eth2]:eth1
Select member priority:
One of [none primary secondary] (default – none):
Enter member name, <Enter> to end:
One of [eth0 eth2]:eth2
Select member priority:
One of [none primary secondary] (default – none):
Enter member name, <Enter> to end:
One of [eth0]:
//选择组成员 eth1,eth2,并直接回车设置优先级为无,最后直接回车结束选择成员
Enter virtual adapter name: vnet0
//将虚拟网络适配器命名为 vnet0
team ioteam activated
//这样 ioteam 就被激活。我们可以通过 View team 来查看已经建立的组:
One of [ioteam]: ioteam
Mux　　ioteam
Mux State　　Active
Teaming Mode　　ALB
Current Primary　　eth2
RLB State　　enabled
Forward Delay　　0
Probes State　　disabled
Mux Speed　　2000
VLAN State　　Disabled
Member　eth1
　　Initiator　　YES
　　Speed　　1000
　　Link Status　　up
　　Priority　　no priority
　　State　　Active
Member　eth2
　　Initiator　　YES
　　Speed　　1000
　　Link Status　　up
　　Priority　　no priority
　　State　　Active
Vadapter　　vnet0
//最后选择 5 存盘,然后选 6 退出即可
①Add team
②Del team
③View team
④Configure team
⑤Save topology
⑥Exit
Select menu option: 5
Enter file name to save to (default – /etc/ians/ians.conf):
Do you want to save the topology across boots? ([Y]/n): Y

完成后系统在/etc/sysconfig/network – scripts 目录下生成了名为 ifcfg – vnet0 的配置文件,这即是生成的虚拟网络适配器的配置文件。在笔者的实例中配置如下:

DEVICE = vnet0
ONBOOT = yes
BOOTPROTO = static

IPADDR = 192.168.77.10

NETMASK = 255.255.0.0

这样就完成了 iANS 的配置。

3 性能测试

网络适配器分组后，需要对服务器网络实际性能进行测试，并依据测试数据判断 iANS 是否能满足我们的需求。为使测试更加符合应用环境，我们采用 FTP 为手段测试传送 360M 数据所耗费的时间，以此估计其带宽容量。

3.1 硬件环境

一台已经安装配置好 iANS 的 I/O 服务器：CPU 为双至强 2.8G，4GB 内存，双千兆网络适配器。若干台计算节点：CPU 为双至强 2.8G，2GB 内存，单千兆网络适配器。交换机为一台 Cisco4506。

3.2 网络性能评价标准

通过 iANS 的安装，我们已经将 I/O 服务器的 2 块千兆网络适配器以 ALB 的模式进行了分组。对于网络而言，其性能主要有以下 5 项指标：

a. 可用性（Availability）

b. 响应时间（Response Time）

c. 网络利用率（Network Utilization）

d. 网络吞吐量（Network Throughput）

e. 网络带宽容量（Network Bandwidth Capacity）

其中可用性往往是确定网络是否正常工作的第一步骤，最简单的办法就是使用 ping 命令发送 icmp 包，并等待另一方返回信息。这项性能一般容易满足。而在本文中，我们最关心的是网络带宽容量（也即能达到的最大带宽）是否能达到我们所预期的水平。

3.3 网络性能测试结果

通过一台计算节点与一台 IO 节点进行数据传输以及两台计算节点同时与 IO 节点进行数据传输，粗略测试分组技术对数据传输的影响，测试结果见下表：

表 1　单计算节点单 IO

测试编号	1	2	3	4	5	均值
时间（s）	4.2	4.5	4.3	4.4	4.5	4.38
速率（MB/s）	85.7	80.0	83.7	81.8	80.0	82.2

表 2　两计算节点单 IO

测试编号		1	2	3	4	5	均值
节点 1	时间（s）	4.5	4.6	4.5	4.4	4.3	4.46
	速率（MB/s）	80.0	78.3	80.0	81.8	83.7	80.7
节点 2	时间（s）	4.4	4.3	4.3	4.6	4.5	4.42
	速率（MB/s）	81.8	83.7	83.7	78.3	80.0	81.4

由于传输数据时帧头等因素的影响，实际传输数据量（包括控制信息）要比原始数据大（约为原始数据的 1.1－1.2 倍）。除去帧头和交换机非线性等影响，仍能达到 65% 以上的利用率，根据笔者的经验，已经比较理想。

4 结论

对于 Cluster 系统中 I/O 节点，我们用多块网络适配器同时为其提供服务，使其传输容量融合在一起，提高了系统 I/O 节点的性能，同时依靠冗余保证网络的畅通，提高了系统的可靠性。实验数据表明，使用 iANS 技术可以方便的达到提高 Cluster 系统可用性（Availability）和提高网络带宽容量（Network Bandwidth Capacity）的要求，具有效率高，对上层用户透明，以及较好的负载均衡性等优点，并且不需要专用交换机支持。当集群中 I/O 节点网络性能不能很好满足需求时，可以通过增加网络适配器及线路条数，实现 I/O 节点带宽的线性增加，提高整个系统性能。

参考文献

1 赵改善，包红林. 集群技术及其在石油工业的应用. 石油物探，2001，40(3)：118－126.

2 White Paper，Intel Advanced Network Services software，Support. intel. com.

3 胡修林，王运鹏，郭辉. 多网卡链路绑定策略的研究与实现. 小型微型计算机系统，2005，26(2)：165－168.

4 肖文名，李永生，陈晓宇，宗翔. 高性能计算系统性能评测关键问题探讨. 计算机系统应用，2008，17(3)：115－118.

5 屈刚，邓健青，韩云路. Linux 集群技术研究. 计算机应用研究，2005，(5)：100－101，107.