面向地震资料处理的集群系统设计与实现①

金 弟 薛中州 杨 俊 (中国石油杭州地质研究院 计算机应用研究所 浙江 杭州 310023)

摘 要: 针对能源应用领域的地震资料处理对高性能并行计算的需求,介绍了一种集群系统的设计与实现方案。对集群系统中易产生性能瓶颈的网络通信子系统、存储子系统、全局共享文件系统这些关键子系统提出了详细的设计方法与实现技术。同时对集群系统管理、地震资料处理的并行计算应用部署给出了具体的实现方法。最后分别使用 LINPACK 基准测试与地震资料处理并行计算应用实测结果,验证该集群系统在高性能并行计算方面的优越性。

关键词: 高性能计算; 集群系统; 地震资料处理

Design and Implementation of Seismic Data Processing Oriented Cluster System

JIN Di, XUE Zhong-Zhou, YANG Jun

(Department of Computer Application, Research Institute of Hangzhou Geology, Hangzhou 310023, China)

Abstract: To meet the requirement of good performance of parallel computing in the field of energy application of seismic data processing, this paper expounds the design and implementation of cluster system. For the key subsystems tending to produce performance bottleneck like network communicaton subsystem, storage subsystem and global share file system subsystem, the article proposes designing and implementation technology in detail. The cluster system management and parallel computing application of seismic data processing deploying are introduced. Finally, testing in linkpack benchmark and seismic data processing application prove the system's advantages in parallel computing performance.

Keywords: high performance computing; cluster system; seismic data processing

1 引言

能源应用领域的地震资料处理是高性能计算的主要应用之一, 2008 年中国高性能计算 TOP100 中, 能源应用领域占了 35%[1],遥遥领先于其他应用领域。集群技术是最近几年迅速发展的一项高性能计算技术, 它由一组相互独立的计算机通过高速的通信网络互连而组成,并以单一系统的模式加以管理。具有低成本、高性能、高扩展性、高吞吐量和易用性等特点[2]。在 2008 年全球高性能计算 TOP500 中, 采用集群技术占 74.6%[3]。

本文讨论的集群系统是面向石油行业的地震资料处理应用领域而设计的基于 Linux/Unix 混合的计算机集群系统。目前该系统已经广泛应用于陆地及海上的二维、三维地震资料的叠前偏移, 表现出整体性能好、可靠性高、可扩展性强以及较高的性价比。

2 集群系统设计与实现

2.1 总体结构

该集群系统的总体结构如图 1。主要由 64 个计算节点,4 个 I/O 节点,1 个管理节点、1 套 Ciscio6506 千兆交换机、1 套 McData4700 光交换机、1 套 8TB 容量的 IBM DS4800 存储、1 套 SUNV440 服务器、1 套 3590H 带库等通过百兆管理网、千兆数据与计算网及 4G 光纤 SAN 存储网构成。

2.2 网络子系统设计与实现

网络子系统是关键的组成部分之一, 也是整个系统的瓶颈之一。针对地震资料处理的高吞吐量、高密度数据通信等特点, 不同业务网络流量的规模及集群

① 收稿时间:2009-09-01;收到修改稿时间:2009-10-23

系统的日常远程管理维护等网络需求。在网络设计上采用分类的思想分为物理上独立的三类(链接方式见图 1)：①面向系统管理人员的百兆管理网络，主要用于对整个集群系统的安装、调试、故障诊断及日常维护管理等。②面向应用处理的千兆数据与计算网络，主要用于地震数据的分发、传输收集及并行计算处理等。③面向 I/O 读写的 4G 光纤网络，主要用于大规模 I/O 数据的读写等。分类的网络设计保证了不同业务同时运行时彼此不受影响、某一网络故障局部受影响、网络性能按需分配等优点。

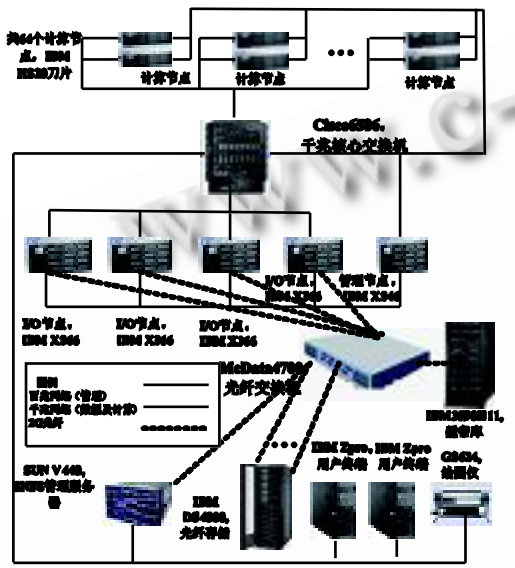

图 1　集群系统总体结构设计

在网络技术实现上主要采用了 Vlan 技术、Zone 技术以及网卡 Bonding 技术。Vlan 技术将一个物理的 LAN 逻辑地划分成不同的广播域，与物理上形成的 LAN 有着相同的属性,从而有助于控制流量、隔离广播风暴，减少设备投资、简化网络管理、提高网络的安全性。在千兆数据与计算网络中，把用于业务数据与 SNFS 文件系统控制信息数据划分不同的 Vlan。用于业务数据 Vlan 主要配置如下：

进入 CiscoCatalyst6506 特权模式，进行如下配置。

```
Conf  t
interface Vlan1
ip address 192.168.8.252  255.255.255.0
ip helper-address 10.183.1.1
ip helper-address 10.183.1.4
ip helper-address 10.183.1.5
no ip proxy-arp
```

Zone 技术使连接在 SAN 网络中的设备，逻辑上划分为不同的区域，可以提高安全性、负载均衡以及操作方便。在4G光纤网络中把 4个I/O节点、DS4800 存储及 SNFS 管理服务器 SAN 设备划分成 2 个 Zone，使得 DS4800 存储的 4 个奇数端口与 4 个偶数端口分别属于不同的 Zone。主要配置如下：

进入 McData4700 的 Zone 配置模式，采用“域，端口”方法主要实现如下：

```
zonecreate " zone_a", "1,2; 1,4 ;1,6; 1,8 ; 1,9 ;1,10; 1,11; 1,12 ;1,13"
zonecreate " zone_b", "1,1; 1,3;1,5; 1,7 ; 1,9 ;1,10; 1,11; 1,12 ;1,13"
#  9,10,11,12,13为 4个I/O节点与 1个SNFS服务
# 器节点
```

Bonding 技术可以利用高性价比的网络部件，有效提高通信带宽，通过把多块网卡配置绑定到一个 IP 地址，理论上总的带宽可达到单网卡带宽的 n 倍(n 块网卡绑定)。在 I/O 节点采用了三路 Bonding 技术。主要配置如下：

(1) 配置文件 /etc/modeprobe.conf, 内容如下：

```
alias  bondo binding  #加载网卡 bond0 设备模块
```

(2) 配置/etc/sysconfig/network-script 目录下的文件

(a) ifcfg-bond0 内容如下：

```
DEVICE=bond0
IPADDR=192.168.8.21
NETMASK=255.255.255.0
ONBOOT=yes
BOOTPROTO=none
USERCTL=no
```

(b) ifcfg-eth1,ifcfg-eth2,ifcfg-eth3 内容如下：

```
Master=bond0
Slave=yes
```

(3) 配置文件 /etc/rc.local 内容如下

```
ifconfig  bond0  192.168.8.22  netmask
```

255.255.255.0 up

ifenslave bond0 eth1　#网卡 1

ifenslave bond0 eth2　#网卡 2

ifenslave bond0 eth3　#网卡 2

2.3 存储子系统与文件系统

存储子系统与文件系统是影响集群系统整体性能的关键因素，是高性能计算 I/O 的瓶颈。针对地震资料处理的高密度 I/O 应用特点，对存储子系统设计使用 IBM DS4800 作为物理存储设备，采用基于热备盘的 RAID5 模式与双控制器交叉均衡的连接方式，确保存储子系统物理的可靠性、I/O 并行读写及负载均衡性。对文件系统的设计采用了基于 SNFS 文件系统作为集群系统的全局文件系统共享的设计方法。SNFS 是一种可安装、可动态扩充，可动态调整参数的共享并行文件系统，有效实现并发 I/O[4]。该文件系统将实现存储资源共享；利用 I/O 节点的能力实现负载分流、支持更多作业同时进行；通过资源调配、参数调优、可提高存储系统使用率和性能。

(1) 存储子系统设计与实现

整个存储子系统物理上由 28 块 300GB、10000 RPM 高速光纤盘组成，分为 3 组(编号分别为图 1 中的 1，2，3)。存储物理设计如图 2，RAID5 设计如表 1。

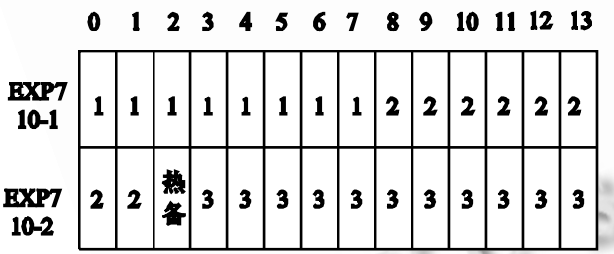

图 2　存储物理设计

表 1　存储系统 RAID 设计

ID	级别	数量	校验	空间
1	5	8	7D+1P	2100G
2	5	8	7D+1P	2100G
3	5	11	10D+1P	3000G

(2) 文件系统设计与实现

对 3 组 RAID 的存储子系统的 LUN 设计如表 2。使用上述的 LUN 参数，实现 SNFS 文件系统设计、实现方法如下：

表 2　存储子系统 LUN 设计

ID	LUN ID	容量 GB	控制器
1	001	525	0
1	002	525	1
1	003	525	0
1	004	525	1
2	005	525	0
2	006	525	1
2	007	525	0
2	008	525	1
3	009	750	0
3	010	750	1
3	011	750	0
3	012	750	1

(a) 在 MetaData Server 服务器(Sun V440 SNFS 管理服务器)上：

Step1：生成 LUN 配置文件，cvlabel -c >/usr/cvfs/config/cvlabels

Step2：编辑生成的 LUN 配置文件，根据文件系统名字修改 LUN 标签。

Step3：LUN 打标签操作，cvlabel /usr/cvfs/cvlabels

Step4：SNFS 文件系统配置文件的设计与优化。根据存储的划分与具体应用，配置优化文件系统相关参数形成 cfg 文件，存放目录/usr/cvfs/config/。

Step5：编辑/usr/cvfs/config/fsnameservers 加入 MetaData Server 服务器 IP；编辑/etc/cvfs/config/fsmlist 加入文件系统名称。

(b) 在 DiskProxy 服务器上(4 个 I/O 节点)上

Step1：编辑/usr/cvfs/config/fsnameservers 加入 MetaData Server 服务器 IP

Step2：使用 sndpscfg -e 生成文件/usr/cvfs/config/dpserver，根据需求进行参数修改

Step3：编辑/etc/fstab 文件，mount 文件系统

hzdata1　/hzdata1 cvfs diskproxy=server 0 0

hzdata2　/hzdata2 cvfs diskproxy=server 0 0

hzdata3　/hzdata3 cvfs diskproxy=server 0 0

(c) 在计算接点上(64 个计算节点)

Step1：编辑/usr/cvfs/config/ fsnameservers,加入MetaData Server 服务器 IP。

Step2：编辑/etc/fstab 文件，mount 文件系统

hzdata1　/hzdata1 cvfs diskproxy=client 0 0

hzdata2　/hzdata2 cvfs diskproxy=client 0 0

hzdata3　/hzdata3 cvfs diskproxy=client 0 0

对上述设计的 SNFS 文件系统与常规的 NFS 文件系统在计算节点上进行了 dd 命令的文件系统读写测试 10 次，取平均值测试结果如表 3。从表 3 的数据看出，使用上述的方法对存储子系统与文件系统的合理优化设计与配置，SNFS 比常规的 NFS 方法在性能上有了大幅度的提升，特别在数据写操作方面效率提高了 45.55%。

表 3　dd 命令对两种文件系统的读写测试结果

类型	数据量 (M)	时间 (秒)	速度 (M/秒)	理论速度 (M/秒)	效率 (%)
NFS 读	1024	9.58	106.89	125	85.51
SNFS 读	1024	9.23	110.94	125	88.75
NFS 写	1024	19.56	52.35	125	41.88
SNFS 写	1024	9.37	109.28	125	87.43

2.4 集群系统管理

集群系统由大量节点组成，如何正确、高效及统一的部署、维护、管理这些节点是集群系统管理的主要内容。该集群系统采用的主要系统管理方法、技术如表 4。

表 4　系统管理方法、技术

管理类型	采用的方法、技术	主要优点
计算节点管理	采用开源 Xcat,集群管理软件（部署在管理节点），通过主要对 passwd. tab、nodetype. tab、site. tab、等配置表合理优化配置、定制满足该集群的实际需求。	实现统一管理各个计算节点，提高管理效率
用户帐号管理	基于全局共享文件系统 SNFS：采用 NIS 技术，管理节点为 NIS 服务器，其他节点为 NIS 客户端。	确保各节点用户环境一致性，易于统一维护与管理帐号
部署管理	采用基于网卡 PXE 协议的 DHCP 进行网络映像部署；	快速部署大规模计算节点，扩展性好、部署自动化
登录管理	采用单一登录的方法，通过开启 rsh,rlogin,rsmo 方法来构建节点间可信任关系。	统一登录、操作方便、易于并行处理

2.5 并行计算应用部署

该集群系统的的应用领域是面向石油地球物理勘探的地震资料处理，上述完成的集群系统的硬件与软件部分的设计与实现主要为应用提供基础设施平台。部署在该集群系统地震资料处理大型并行计算软件分别为美国 WestGeo 的 Omega1.8.3，并行计算环境采用 PVM[5]；法国 CGG 公司的 GeoCluster3.1,并行计算环境采用 MPI。本文以 GeoCluster 为例，规划部署如下：

(1) GeoCluster 的管理用户与普通用户采用 NIS 方式构建，主目录为：/hzdata1/home/cgg

(2) GeoCluster3.1 软件部署分布：

/hzdata1/soft/gct3110

/hzdata1/soft/intel

/hzdata1/soft/oracle

(3) GeoCluster 的 LOGGER 机与许可管理机使用管理节点；

为了对提交的作业进行分发并行计算以及把并行计算结果进行收集，在每个节点的/etc/rc.local 启动如下进程进行计算节点间的协同工作：

/hzdata1/soft/gct3110/jobmgr/admin/gvrstart LOGGER=hzm01 gvr_root=/cgg SITE=HZSOU

3　集群系统整体性能测试

从基准测试与并行计算应用 2 个方面对集群系统的整体性能进行了测试。

3.1 基准测试

采用国际上标准的用于高性能计算机系统浮点性能测试的 LINPACK[6]。本次测试采用该集群系统的 64 个节点，128 个主频 3.4GHz Xeon CPU,内存共 256GB(每个计算节点 4G)，RedHatAS3U5 64 位操作系统。测试软件为：GNU 编辑器 gcd，MPICH-1.2.7.tar.z，ATLAS.tar.z,HPL.tar.z。测试命令为 mpirun－np 128 xhpl。通过 10 次测试取平均值得到的测试性能参数如表 5。

表 5　基准测试数据

CPU	主频	理论峰值	实际峰值	性能比例
128	3.4Ghz	870.4 亿次	467.4 亿次	53.6%

表 6　2008 年中国 TOP100 基准测试数据

年份	机器类别	性能比例
2007	国内机器	33.1%
	国外机器	66.9%
2008	国内机器	55.93%
	国外机器	44.07%

通过表 5 与表 6 可知，从 Linpack 测试的角度来看集群系统的整体性能比例，本文给出的集群系统的性能比例比 2007 年 TOP100 中的国内机器高 20.5%，比 2008 年的国外机器高 9.53%，集群系统整体性能表现良好。

3.2 并行计算应用测试

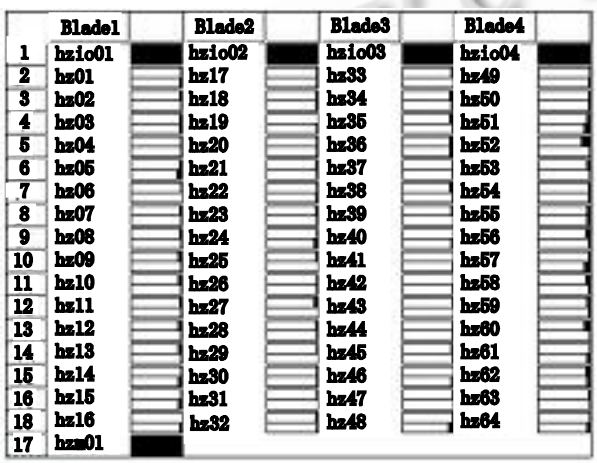

图 3 并行计算应用软件实测结果

采用 GeoCluster3.1 地震资料处理软件进行测试。测试环境：采用中国某海洋区域的实际一束 20 条测线，总共约 4G 数据规模进行叠前时间偏移，采该集群系统的全部 64 个节算节点、128 个 CPU 进行并行计算，整个集群系统 CPU 资源使用效率如图 3。图中 Blade1 至 Blade5 为 5 组刀片，hzio01 至 hzio04 为 4 个 I/O 节点，hzm01 为管理节点。hz01-hz64 为 64 个计算节点，其中白色部分代表 cpu 处于用户模式，黑色部分代表 cpu 处于核心模式。由图可知，该集群系统在进行地震资料偏移并行计算时，64 个计算节点的 CPU 负载均衡，基本上 100% 高负荷的使用效率，利用效率都非常高，而且大部分处于用户模式运转，没有出现由于网络传输、I/O 等待的瓶颈现象，该集群系统在面向地震资料处理的高性能计算应用方面，各子系统协同工作，整体性能表现良好。

4 总结

根据地震资料处理对高性能并行计算的需求，本文介绍了一种集群系统的设计与实现方案。对网络通信子系统、存储子系统、文件系统、集群系统管理、地震资料处理的并行计算应用部署进行了详细的设计与实现。最后使用 LINPACK 基准测试与在地震资料处理中的应用测试，验证该集群系统在高性能并行计算方面的优越性。

参考文献

1 张云泉.2008 年中国高性能计算机TOP100 排行榜分析与展望.北京.2008 年全国高性能算法软件研究开发研讨会, 2008.

2 Buyya R. High Performance Cluster Computing Architectures and Systems, Volume 1. Posts&Telecommunications Press. 2002.

3 电脑商网.刀片＋集群:高性能计算新潮流.电脑商报, 2008,3:031－031.

4 Quantum coporation. StorNext File System. ADIC Educational Services. 2006.

5 Buyya R. Programmingand Applications, Volume2. Posts&Telecommunications. Press. 2002.

6 Linpack ome. Page. [2009-4-10] http://www.netlib.org/linpack