基于MPICH2的高性能计算集群系统研究

王勇超[1],张 璟[1],王新卫[1],马 静[2]
(1.西安理工大学,陕西 西安 710048;
2.西安工业大学,陕西 西安 710032)

摘 要:目前在高等学校和科研机构中对于高性能计算的需求很大,而商业的超级计算机性能虽高但价格昂贵,同时这些单位又都拥有大量普通的PC机和网络设备。为了利用现有硬件资源获取高性能计算能力,文中研究了在PC机和Linux环境下构建基于MPICH2的高性能计算集群系统的方法,搭建了一个拥有16个节点的系统并利用高性能Linpack基准测试方法进行了性能测试。测试结果表明,这种构建高性能计算集群系统的方法切实可行,是低成本获取高性能计算能力的良好途径。

关键词:高性能计算;集群;PC机;并行计算;Linux;MPICH2
中图分类号:TP302 **文献标识码**:A **文章编号**:1673-629X(2008)09-0101-04

Research of High Performance Cluster System Based on MPICH2

WANG Yong-chao[1],ZHANG Jing[1],WANG Xin-wei[1],MA Jing[2]
(1.Xi'an University of Technology,Xi'an 710048,China;
2.Xi'an Technological University, Xi'an 710032,China)

Abstract:High performance computation is in great demand in universities and research institutions these days. The performance of commercial supercomputer is very high, but it is very expensive. At the same time, there are a mass of ordinary PC and network equipment in university and research institutions. In order to take advantage of existing hardware resources access to high-performance computing capabilities, researched the method in building up a high performance computer cluster system which was based on MPICH2 in Linux operating system on PCs and built a system which has sixteen nodes. Meanwhile HPL testing was applied to measure system performance And the testing results made it clear that this method was not only very useful in building up cluster systems but also cost-effective.

Key words:high performance computation;cluster;PC;paralleled computing;Linux;MPICH2

0 引 言

随着科学技术的发展,大型科学与工程计算对计算性能要求不断提高。但是限于处理器制造工艺,单颗处理器性能可提高的空间已越来越小,在这种情况下,高性能并行计算越来越受到人们的重视和青睐。在2006年6月世界超级计算机Top500排行中,有364台采用了集群体系,而五年前,这一数字仅仅为32台。集群已经成为超级计算机的发展趋势。

IBM、HP等著名计算机公司研制开发的集群超级计算机性能虽高,但价格也非常昂贵。所以,研究使用廉价PC机在局域网内构建性能稳定而且功能强大的高性能计算集群具有十分重要的实际意义。

文中将研究利用PC机,在Linux环境下构建基于MPICH2的高性能计算集群系统的方法,并对集群系统进行高性能Linpack测试。

1 集群系统及并行计算原理介绍

1.1 集群系统原理

将多台同构或异构计算机连接起来协同完成特定任务就构成了集群系统[1]。集群系统主要分为两种:高可用性集群和高性能集群。高可用性集群主要功能就是提供不间断服务,适用于必须一天二十四小时不停运转的计算机环境。高性能集群是通过将多台机器连接起来同时处理复杂计算问题,应用在需要大规模科学计算的环境中,如天气预报、石油勘探、分子模拟、基因测序等。

1.2 并行计算和MPICH2

高性能计算集群实际上是通过并行计算来实现计

收稿日期:2007-12-10
基金项目:陕西省科技计划项目(2006K04-G10)
作者简介:王勇超(1979-),男,河北定州人,助教,硕士,研究方向为高性能计算、计算机应用;张 璟,教授,博士生导师,研究方向为计算机网络、软件开发以及电子商务。

算性能的提高,并行计算[2]的基本思想是多个处理器协同求解同一问题,即将被求解问题分解成若干个部分,各部分均由一个独立处理机来并行计算。并行计算的优点是具有巨大的数值计算和数据处理能力。并行计算系统既可以是专门设计、含有多个处理器的超级计算机,也可以是若干台以某种方式互连的独立计算机构成的集群[3]。文中搭建的集群系统就是通过以太网方式互连的多台独立 PC 机。

目前两种最重要的并行编程模型是数据并行和消息传递。数据并行指将相同操作同时作用于不同数据,从而提高问题求解速度。数据并行编程模型的编程级别较高,编程相对简单,但只适用于解决数据并行问题。消息传递模型各个并行执行任务之间通过传递消息来交换信息、协调步伐、控制执行。消息传递模型编程级别较低,编程相对复杂,但为程序员提供了更加灵活的控制手段和表达形式,可以实现一些用数据并行模型很难表达的并行算法[4]。文中采用的就是消息传递模型。

基于消息传递并行编程模型的并行编程语言主要有 PVM(并行虚拟处理机)和 MPI(消息传递接口)两种。MPI 吸收了现存的许多系统的最突出优点,是当今最为流行的用于并行编程的消息传递库标准[5]。MPI 有很多具体实现,其中 MPICH 是 Linux 平台下最重要的一种 MPI 实现。MPICH 是一个与 MPI 规范同步发展的版本。每当 MPI 标准推出新的版本时,MPICH 就会有相应的实现版本,目前的最新版本是 MPI-2。MPICH2 是一个全面支持 MPI-2 标准的 MPI 实现。与之前版本相比 MPICH2 具备更加严谨和合理的结构,可移植性和效率更好[6]。文中采用的就是 MPICH2 1.0.3。

2 集群系统的构建

2.1 集群系统硬件环境

构建集群系统,首先要将所有计算机配置在同一个局域网内。笔者所搭建的集群系统共有 16 台 PC 机,系统硬件环境结构如图 1 所示。每个节点 PC 机配置为:INTEL P4 2.0G CPU,512MB 内存,40G 硬盘,一块 100Mb/s 网卡。搭建集群的硬件环境就是利用超五类双绞线通过交换机把 16 台 PC 机组成一个局域网。

2.2 集群系统的软件环境

2.2.1 操作系统

Unix 在超级计算机操作系统领域占统治地位。Unix 运行稳定、安全性也比较好。基于 Unix 的开源免费 Linux 自 20 世纪 90 年代末以来不断走向成熟,健壮性不断增强,并且提供了 GNU 软件和标准化的 PVM、MPI 消息传递机制,提供了对高性能网络支持。2006 年 6 月超级计算机 TOP500 中 73.4% 采用 Linux。目前 Windows 在个人操作系统中采用较多,且多数并行计算环境也支持 Window,所以也可以在采用 Windows 作为集群操作系统[7]。笔者所搭建集群采用的操作系统是 RedHat Linux 9.0。

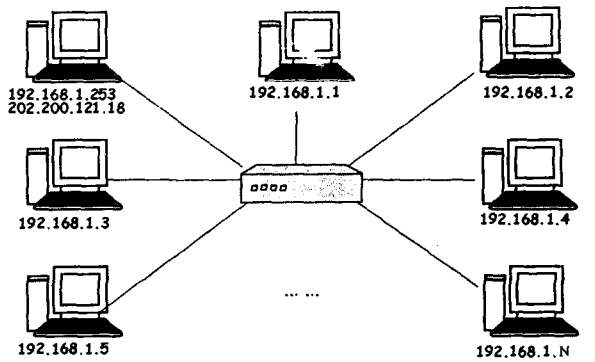

图 1 集群系统硬件环境结构图

2.2.2 网络配置

集群内节点的主机名按照顺序统一命名,如“node01,node02…”,安装上 TCP/IP 协议,统一分配 IP 地址,如:192.168.1.1,192.168.1.2…。其中一台 PC 机作为主节点,主机名为“node0”,内网 IP 地址“192.168.1.253”。同时主节点作为与外部网络的接口,分配外网 IP 地址“202.200.121.18”。

2.2.3 单一登录构建

所谓的单一登录最直接的外在表现就是,无论是用户提交任务还是管理员进行管理操作,登录集群系统只需要输入一次用户名和密码就可以操控集群系统中的任意一个节点。

在集群系统中,构建单一登录可以通过远程通信协议来完成。当前用于远程通信的协议很多,其中比较出名的是 RSH 和 SSH(Secure Shell)[8]。SSH 是 RSH 的一种改进,SSH 是通过 SSL 的加密方式来传输数据,从而避免了数据被截获和修改的可能。SSH 还可以实现远程登录,任何节点都可以通过 SSH 登录到其他节点进行权限允许内的操作。有了 SSH,只需要登录主节点,就可以控制集群内的所有节点,从而实现单一登录。

2.2.4 并行环境构建

MPICH2 的源码可以从网上下载,下载下来之后,首先以 root 用户登录主节点,建立安装目录,如/usr/MPICH2-install,将下载文件解压缩,通过 configure 脚本完成初始配置,然后使用 make 编译,编译成功之后 make install 安装。

安装之后,还需进行配置。首先在 root 目录下编

辑.bashrc文件修改环境变量,即将MPICH2的安装路径加入到文件中,如:PATH = "$PATH:/usr/MPICH2-install/bin"。然后修改/etc/mpd.conf文件,修改内容为:secretword=myword,并设置文件读取权限和修改时间:

```
# touch /etc/mpd.conf
# chmod 600 /etc/mpd.conf
```

最后在root目录下建立节点文件mpd.hosts,内容为所有节点的主机名,如:

```
node01
node02
.....
node15
```

安装配置完成之后,主节点的并行环境就构建好了,每一个参与计算的节点都必须安装MPICH2。

2.2.5　单一文件系统构建

由于集群系统是多节点协同工作,大量相同的软件和数据要安装在所有的节点上,同时大量的数据共享也是必不可少的。比如安装并行环境MPICH2,若构建集群并行环境,必须在所有的节点上都安装上MPICH2,集群系统有多少个节点就得安装多少遍。

为了解决这种问题,就必须建立单一文件系统。NFS(NetWork File System)网络文件系统是集群系统中解决这个问题的一个非常有效的方法。NFS是一种使用比较广泛的网络文件系统。将主节点配置为NFS服务器,子节点通过NFS服务可以把需要共享的文件、文件夹或者分区共享给网络上的其他计算机,需要访问这些数据的计算机使用mount命令把共享的资源加载到自己的系统上,然后就可以像使用本地文件系统一样方便。为了方便可以修改/etc/fstab文件,使子节点开机即自动加载。

3　系统性能测试与分析

3.1　HPL测试介绍

Linpack是通过对稠密线性代数方程组求解能力的测试评价高性能计算机系统的浮点运算性能,测试结果按每秒浮点运算次数(Flop/S)表示。HPL(high performance Linpack)是第1个标准的公开版本并行Linpack,$N \times N$测试的MPI实现,可适应多体系移植,目前广泛用于Top500测试。这一测试主要针对分布式存储大规模并行计算系统而设计,用户可以设置任意大小的问题空间,使用任意个数的CPU,利用基于高斯消去的各种优化方法寻求最佳的测试结果。

3.2　HPM输入参数

HPL测试中输入参数的设置对测试结果影响很大。HPL的输入参数在HPL.dat文件中设置。可以设置的参数有很多,对性能测试结果影响较大的主要有问题空间的大小(Ns)、LU分解数据块大小(NBs)。

3.3　Ns对HPL测试的影响

问题空间(Ns)即集群求解问题的规模,其最大取值与可用内存大小有关。HPL测试程序计算使用64位精度的矩阵,所以设N为问题空间,M为内存总量(单位MB),则N与M之间存在如下关系:

$$N^2 \times 64 = M \times 1024 \times 1024 \times 8$$

即:

$$N = \sqrt{M \times 1024 \times 1024 \times 8/64} \approx 326\sqrt{M}$$

据此可计算问题空间最大值,大于这个值,计算时内存不足系统将使用交换分区,导致计算速度迅速降低。在实际当中,由于系统及计算本身要占用部分内存,所以一般取可用内存的80%。假设集群共有P个计算节点,所有节点可用内存均为m(单位MB),则问题空间取值如下:

$$N = 326\sqrt{m \times 0.8 \times P}$$

保持其他参数不变,不断调整问题空间,测试结果如图2所示。

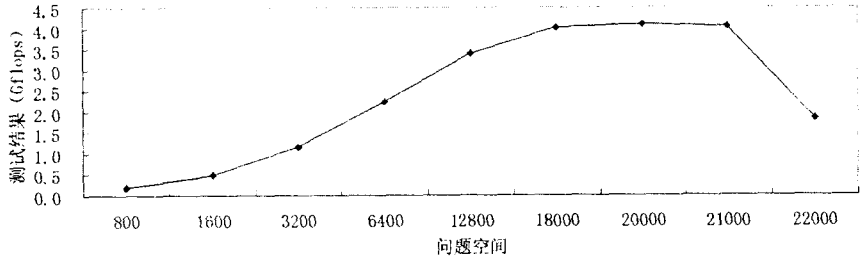

图2　问题空间对测试结果的影响

从图看出,同样的配置,随问题空间的增大,测试结果迅速增高,达到一定高度后又迅速降低。问题空间为800时,实测结果为0.1941GFlops,不但不及问题空间为20000时的5%,且远低于单机测试结果最好值的0.37GFlops。原因是问题空间太小,网络通信等额外开销远远超出了节点增加带来的性能提高。

3.4　NBs对HPL测试的影响

LU分解数据块大小NBs为矩阵分解为小数据矩阵的大小。NBs的最佳取值没有明确依据,主要在具体环境下通过多次实验不断摸索尝试。网络结构、问题空间、集群计算节点数量等诸多因素对NBs取值都有影响。

首先,在单节点下在不同问题空间下测试NBs对测试结果的影响。测试结果如图3所示。

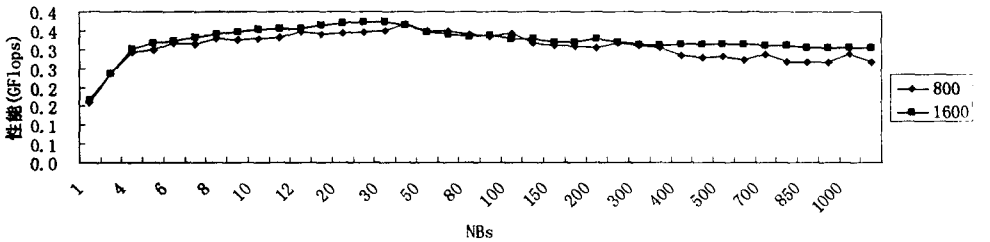

图3 测试NBs对测试结果的影响

从图3中可以看出,在单节点下,NBs对测试结果的影响并不是很明显。

分别在1个节点,4个节点,9个节点和16个节点下测试NBs对测试结果的影响,测试结果如图4所示。从图中曲线可以看出,当节点数大于1时,NBs对测试结果影响比较明显,原因是NBs取值影响负载均衡。NBs取值太小,虽然可以更好实现负载均衡,但不能充分发挥节点CPU处理性能,且计算过程需要不断进行网络通信,这时候网络带宽和延迟对测试结果将产生很大影响。NBs取值过大,导致负载不均衡,根据木桶原理,最终测试结果反而降低。极端情况下,NBs等于或大于Ns,负载极度不均衡,这种情况下测试结果甚至远远低于单个节点。

3.5 节点数对HPL测试的影响

对问题空间分别为3600和7200两个问题分别用1,2,4,6,9,12,15,16个节点进行求解,求解该问题计算耗时对比曲线和计算性能对比曲线如图5所示。

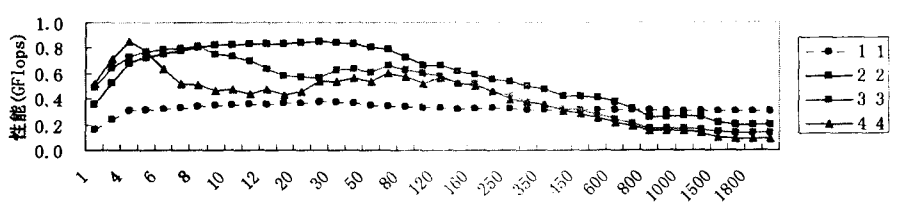

图4 在不同节点下测试分块大小对性能影响测试

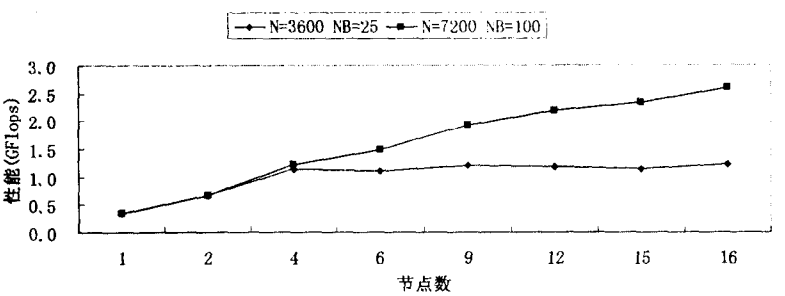

图5 同一问题下不同节点数对计算性能的测试结果

通过图5可以看出,随着计算节点的增加,若问题空间太小,网络通信等额外开销不能抵消计算性能提高带来的好处,则整体计算性能测试值不会提高,反而有可能降低。且在没有超过问题空间最大值的情况下,问题空间越大,测试性能越高。

在实验中进行多次测试,发现当问题空间达到20000时,实测结果为4.10GFlops,超出单节点测试最高值0.37GFlops的11倍还多。由于额外开销的增加并没有达到理想状态下的16倍,但是多节点并行计算的巨大威力已经显示了出来。

4 结束语

文中利用普通PC机在Linux环境下构建了基于MPICH2的高性能计算集群系统,并对其进行了高性能Linpack测试。实验结果表明这种构建高性能计算集群系统的方法切实可行,不但可以节省计算时间,提高计算精度,而且可以解决单机无法完成的超大规模求解问题,是一般单位和科研院所低成本满足超级计算需求的最佳途径。同时通过HPL基准测试,证明影响这种集群系统测试性能的主要参数问题空间的大小(Ns)、LU分解数据块大小(NBs),影响集群系统整体性能的主要因素是计算节点的中央处理器(CPU)数量、内存以及网络架构(主要是网络带宽和延迟)。

参考文献:

[1] 屈 钢,邓健青,韩云路.Linux集群技术研究[J].计算机应用研究,2005(5):100-101.

[2] 陈国良.并行计算—结构、算法、编程[M].北京:高等教育出版社,2004.

[3] 车静光.微机集群组建、优化和管理[M].北京:机械工业出版社,2004.

[4] 孙世新,卢光辉,张 艳,等.并行算法及其应用[M].北京:机械工业出版社,2005.

[5] 宋 伟,宋 玉.基于SMP集群系统的并行编程模式研究与分析[J].计算机技术与发展,2007,17(2):164-167.

[6] 孙亦嘉,张 岳,陈 渝.基于VIA的MPICH2研究与实现[J].计算机工程与应用,2005(1):99-101.

[7] 张翠莲,刘方爱,王亚楠.基于MPI的并行程序设计[J].计算机技术与发展,2006,16(8):72-74.

[8] Buyya R,Cortes T,Jin Hai.Single System Image[J].The International Journal of High Performance Computing Applications,2001,15(2):124-135.