文章编号:1007 – 6492(2001)04 – 0028 – 03

高性能桌面超级并行计算机 Linux PC 集群系统

王文义[1],王若雨[2]

(1.郑州大学信息工程学院,河南 郑州 450002;2.郑州电力职工大学,河南 郑州 450052)

摘 要:并行处理技术是衡量一个国家综合国力的重要标志之一,而桌面超级并行计算机——Linux PC 集群系统由于具有可扩展性好、成本低廉、高性能和能够获得免费的 Linux,PVM 及 MPI 并行编程环境等优点,从而特别适合我国国情.根据当前并行处理硬、软件资源的发展状况,提出了 Ethernet,Fast Ethernet,Myrinet,ATM 和 Giganet 5 种构建 Linux PC 集群系统的选择方案.

关键词:并行处理;Linux PC 集群;性能价格比;浮点运算速度

中图分类号:TP 316.4 文献标识码:A

在国民经济、国防建设和科技发展中存在着许多具有广泛而深刻影响的重大课题,如石油勘探、地震预测、大范围天气预报、人体基因和遗传工程、核武器系统的研究和模拟等,与此相关的具体应用领域,如求解量子力学、聚合化学、晶格生长、流体动力学、量子场理论、分子动力学、使用多级牛顿算法求解大规模电路方程以及进行 VLSI 芯片设计等,均需要由大容量、高速度的计算机系统来完成.显然,传统的 Von Neumann 型单处理计算机已远远不能满足上述各方面对性能的要求.一方面,计算机的速度要受到顺序执行的限制;另一方面,VLSI 器件本身的开关速度也存在一定的极限,更何况电信号的传播速度最快也不能超过光速.因此,开发高性能的超级计算机已成为时代的需要,而并行处理技术则是这种开发工作的唯一选择.

1 并行处理系统

1.1 并行计算机体系结构

高性能计算机通常是指采用各种并行技术的机器,它们可分为 4 类:第一类是多向量计算机,如 CRAY – C90,Fujitsu VP2000 等;第二类是基于共享存储的多处理机系统,如曙光Ⅰ号,HP/Convex Exemplar SPP1200 等;第三类是基于分布存储的 MPP 系统,如 IBM SP2,SGI T3E1200 等;第四类则是基于 RISC 工作站或各种 PC 经由高速互连网连接而成的集群计算机等.其中第一类和第三类系统由于研制费用高及价格昂贵等因素,现已逐步退出了市场,如表 1 所示.

表 1 几种多向量计算机和分布存储型计算机的性能价格

Table 1 The performance/price of some multicomputer and distributed memory computer

机型	体系结构	峰值速度/Gflops	市场售价/百万美元	芯片价格/美元	生产时间
CRAY – C90	多向量计算机	16	30	1875	1996
HP SPP1000	多向量计算机	14	1.7	121.4	1996
IBM SP2	分布式存储	76	2.0	26.3	1999
SGI Origin2000	分布式存储	6.4	0.27	42.2	2000

第二类系统由于受共享结构的限制,使得系统的可扩展性差,规模不可能做得太大,一般只能做到 32 个处理器左右,如 HP/Convex Exemplar SPP1200 就是一个典型,而集群计算机则由于其优越的 3P(Performance Per Price)性能,目前已成为并行处理的热点和主流,尤其是用 PC 构建的桌面集群系统,非常适合于我国国情,它具有硬件资源充沛,价格低廉等突出优点,如表 1 中所列的售价为 27 万美元的 SGI Origin 2000 系统,我们只需花费 20 万元人民币,就可以构造出具有同等性能

收稿日期:2001 – 04 – 10;修订日期:2001 – 06 – 11

基金项目:国家社会科学基金资助项目(97BTQ004)

作者简介:王文义(1947 –),男,河南省洛阳市人,郑州大学教授,主要从事算法分析和并行处理方面的研究.

的 Linux PC 集群计算机，但价格却只有原来的十分之一，因此这类计算机在我国将会有着广阔的市场前景.

1.2　高级处理器芯片和几种新型并行系统

从处理器芯片的研发历史来看，其经历和趋势可表示为：

CISC → VECTOR → RISC → VLIW → Quantum …

根据摩尔定律，芯片制造商大约每 18 个月就能把挤在指甲盖那么大的硅片里的晶体管数目增加一倍，但是这个速度不可能会无止境地持续下去，因为当晶体管间距小到一定程度时，电子运动就会打乱其固有规律而使芯片失效.对此，科学家们的对策是一方面靠提高芯片性能，如目前正在开发研究单机性能高达 100Gflops(每秒 1000 亿次浮点运算)的处理器；另一方面，则通过增加并行系统规模(即增加处理器数目)来提高并行系统的整体性能，如美国 Columbia 大学的 QCDSP 由 20480 个 Pentium 处理器组成；APEnext(欧共体的一种计算机)由 128 个 DEC Alpha21164 VLIW 处理器组成；NERSC IBM SP 由 1024 个 Power3 处理器组成；SGI Blue Mountain 由 6144 个处理器组成；更有甚者，正由 IBM 公司研制的“Blue Gene”(兰色基因)超级计算机将由 100 万个每秒为 10 亿次浮点运算的处理器组成，届时，该计算机的浮点运算速度将达到每秒 1000 万亿次.

2　PC 处理器芯片的性能价格态势

目前高性能处理器芯片主要有 SGI MIPS，HP/PA RISC，DEC Alpha，Sun SuperSPARC，IBM PowerPC，Intel Pentium Pro 等.对于这些处理器，且不说它们的昂贵价格，由其构建的并行计算机，由某些西方国家售往我国时，还要加上诸多不合理的限制，如作为民用目的的计算机，峰速不得超过 12Gflops(120 亿次)；作为军用目的的计算机，峰速则不得超过 7Gflops，这就充分暴露出了其超级大国的霸权本质.

随着现代科技的发展，用于个人计算机的 PC 处理器，其性价比也越来越高，如 Pentium Ⅲ 500MHz 的芯片性能已达到 200Mflops，据今年年初美国市场的统计，各种 PC 处理器芯片的性能价格见表 2.而现在这些处理器售价在国内又下降了许多，看来这种趋势今后还会不断持续下去，这就为我们构建高性能 Linux PC 集群创造了必要的条件.

表 2　各种 PC 处理器芯片的性能价格(2000 年初)

Table 2　The performance/price of various processor chip(at the beginning of the 2000 year)

公司	型号	时钟频率/MHz	价格/美元
Intel	Celeron	700	192
Intel	Celeron	667	170
Intel	Celeron	633	138
Intel	Pentium Ⅲ	600	429
Intel	Pentium Ⅲ	800	729
AMD	K6 – 2	500	69
AMD	K6 – 2	550	89
AMD	K6 – 2	650	180
AMD	K7	650	199.99
AMD	K7	700	234.99
AMD	K7	900	699.99

3　构造廉价高性能 Linux PC 集群的硬软件资源及网络环境需求

3.1　硬件——集群的结点组成与网络选择

对于 PC 集群中的每个结点(一个结点可由一台 PC 或若干台 PC 组成)来说，所需要的硬件资源为：①主板；②内存 64 ~ 128 MB；③L2 Cache；Cache 容量为 256 kB ~ 1 MB；④10 GB 以上硬盘；⑤100 M 自适应快速以太网适配器；⑥显示卡；⑦KVM 设备.

整个系统只需一个监视器，一个鼠标和一个键盘，与 KVM 设备配合即可实现全部桌面控制.

网络环境可以有下面几种选择：

(1) Ethernet：Linux 操作系统；10 MB/s 带宽；网络延迟为 100 ms；PCI 总线；交换机或 Hub.

(2) 快速 Ethernet：Linux 操作系统；100 MB/s 带宽；网络延迟为 80 ms；PCI 总线；交换机或 Hub.

(3) Myrinet：Linux 支持库；1280 MB/s 带宽；网络延迟为 9 ms；PCI 总线；交换机.

(4) ATM：Linux 操作系统加上 AAL * 库；1200MB/s 带宽；网络延迟为 120 ms.

(5) Giganet：Linux 操作系统；1000 MB/s 带宽；网络延迟为 300 ms.

3.2　软件

为了降低 PC 集群的成本，我们应尽量选用免费自由软件，它们很多都可以从网上下载，比如 Linux 操作系统，MS C++ 和 Fortran 90 编译程序以及 MPI(Message Passing Interface)并行程序设计环境等.

MPI 是一个消息传递型并行通信程序设计规范，它是一个消息传递库，其函数可以镶嵌在 C，

C++和 Fortran 等语言中. Linux 支持解释型 MPI，通过网络实现并行过程之间的相互通信，传递的消息可以是指令、数据、同步信号或中断信号. 消息传递的并行编程主要是通过调用 MPI 消息传递库来进行的，它实现了处理器之间的数据交换功能，并提供了并行任务之间的同步和收/发数据的接口. 每个任务的发送操作必需与一个接收任务的接收操作相匹配，反过来也一样，这就完善地提供了同步算法和异步算法的程序实现机制.

MPI 的通信规范主要提供阻塞式(Blocking)和非阻塞式(Non - blocking)通信方式，通信应答关系十分严谨，两种方式都提供点对点通信和聚合通信(Collection). 点对点通信包括 3 种模式：①标准(Standard)模式；②同步(Synchronous)模式；③预备(Ready)模式.

聚合通信包括聚合同步和数据交换两方面. 聚合同步指调用该操作的任务要等到组内所有成员都达到该同步点后才继续往下执行. 数据交换指在一组任务之间一起进行数据交换，它包括 4 种模式：①广播(Broadcast)，组内一个成员的数据发送给所有成员. ②数据分发(Scatter)，组内横向顺序数据交换转为纵向数据. ③数据聚集(Gather)，将一个组内纵向数据交换按序传给各个组. ④全数据聚集(Alltoall)，将所有组内数据进行纵横方向的交换.

4 结束语

使用超级并行计算机开展科学研究，这在过去对于普通科技工作者来说是想都不敢想的事情，但现在它已成为可能. Linux PC 集群将以极为低廉的价格和充沛的硬件资源展现在我们的面前. 如果选择快速以太网互连，那么我们可以花费不到 20 万元人民币，就可构建一台每秒 160 亿次浮点运算的高性能计算机，难道这不是一个惊人的机遇吗？当然，构造 Linux PC 集群毕竟是一项极其复杂的工程，其关键的并行程序设计和通信程序设计难度较高，而这方面的人才在国内几乎处于空白状态. 现在，我们完全可以通过用 PC 来构建高性能并行计算机以创造良好的实践环境和培养环境.

参考文献：

[1] SRINIVAS Aluru，PRABJU G M，JOHN Gustafson. A random - number generator for parallel computers[J]. Parallel Computing，1992，18：839 - 847.

[2] HANK Dietz. Linux Parallel Processing HOWTO [EB/OL]. http://yara.ecn.purdue.edu/pplinux，1998.

[3] RAJKUMAR Buyya. High Performance Cluster Computing：Architectures and Systems[M]. New York：Prentice - Hall，1999.

[4] KAI Hwang. Advanced Computer Architecture[M]. New York：McGraw - Hill Book Co.，1993.

[5] 徐甲同，李学干. 并行处理技术[M]. 西安：西安电子科技大学出版社，1999.

[6] 郑纬民，汤志忠. 计算机系统结构[M]. 北京：清华大学出版社，1998.

Desktop Parallel Supercomputer Linux PC Cluster System with High Performance

WANG Wen - yi[1]，WANG Ruo - yu[2]

(1. College of Information Engineering，Zhengzhou University，Zhengzhou 450002，China；2. Henan University of Electric Power & Workers，Zhengzhou 450052，China)

Abstract：Parallel processing technology is an important mark of judging comprehensive power for a country. Desktop parallel supercomputer Linux PC cluster system has many advantages，for example，good scalability，low cost，high performance and with portable's free Linux，PVM，MPI programming environment etc.，so it is suitable to China. According to the development situations of parallel processing's hardware and software resources，this paper presents five optional schemes for deploying the cluster system，and they are Ethernet，Fast Ethernet，Myrinet，ATM and Giganet.

Key words：parallel processing；Linux PC cluster；performance per price；flop operating speed