CN43-1258/TP
ISSN 1007-130X
计算机工程与科学
COMPUTER ENGINEERING & SCIENCE
2003 年第 25 卷第 6 期
Vol.25,No.6,2003

文章编号:1007-130X(2003)06-0097-04

超级计算机体系结构及应用情况*

The Architectures and Applications of Current Supercomputers

车永刚[1],柳　佳[1],王正华[1],李晓梅[2]
CHE Yong-gang[1],LIU Jia[1],WANG Zheng-hua[1],LI Xiao-mei[2]
(1.并行与分布处理国家重点实验室,湖南 长沙　410073;2.装备指挥技术学院,北京　101416)
(1.National Laboratory for Parallel and Distributed Processing, Changsha 410073;
2.Institute of Equipment and Command Technology, Beijing 101416, China)

摘　要:本文概述了当今超级计算机体系结构类型及特点,考察了国外超级计算机在气象、核模拟、CFD、生物信息学和医学、天体物理学等典型应用领域的应用情况,分析了超级计算机体系结构与应用领域之间的关系。

Abstract: The paper summarizes the architecture types and characteristics of current supercomputers, and investigates the applications of supercommuters in several typical domains such as weather simulation, nuclear simulation, CFD, bioinformatics/medicine, astrophysics in foreign countries. In addition, it also analyzes the relationship between supercomputer architectures and their application domains.

关键词:超级计算机;体系结构;典型应用

Key words: supercomputer; architecture; typical application

中图分类号:TP303　　**文献标识码**:A

1　引言

超级计算机是获得超高运算性能、解决大型科学计算和海量信息处理问题的重要工具,是各国竞争高端计算领先地位的主要领域。目前超级计算机硬件已经能够提供 TFLOPS 计算能力、TB 内存容量、Tbps 网络带宽[1]。

超级计算机性能的提高,主要得益于器件技术、体系结构与互连技术的不断创新与进步。而超级计算的发展,很大程度上依赖于及时开发适用于各种复杂问题数值求解与模拟仿真的应用软件,能否在超级计算机上成功应用,近年来在这方面的进展是巨大的。在气象、核模拟、计算流体力学、生物信息学等领域,超级计算已经成为不可或缺的重要手段[2]。

2　当前超级计算机的体系结构

根据 Top 500 的分类方法,当前超级计算机的体系结构有以下几种。

(1) 对称多处理(SMP)。SMP 结构的计算机

* 收稿日期:2001-10-15;修订日期:2002-05-21
基金项目:国家自然科学基金重点资助项目(69933030)
作者简介:车永刚(1973-),男,云南嵩明人,博士生,研究方向为计算机体系结构与并行计算;柳佳,本科生,研究方向为高性能并行计算;王正华,教授,博士生导师,研究方向为计算机体系结构与并行计算;李晓梅,教授,博士生导师,研究方向为计算机体系结构与并行计算、科学计算可视化。
通讯地址:410073 湖南省长沙市砚瓦池正街 47 号并行与分布处理国家重点实验室
Address: National Laboratory for Parallel and Distributed Processing, 47 Yanwachi St, Changsha, Hunan 410073, P.R.China

一般在单个机柜中包含两个以上处理器，各处理器完全相同，平等地访问软硬件资源，处理器间通过总线或者交叉开关相连，共享存储器，但有各自独立的 Cache。SMP 的优势在于其透明的编程模式，串行程序一般可不加修改直接运行于 SMP 之上。缺点是由于对公共内存和 I/O 的竞争，加上维护 Cache 一致性的开销，导致扩展能力有限。

(2) 大规模并行处理(MPP)。MPP 指在同一地点由大量处理器构成的并行计算机，一般以通用 64 位微处理器作为处理节点，多为分布存储方式，节点间通信用消息传递方式，其规模可扩展到数千节点。MPP 系统的优点是峰值速度高，并有良好的可扩展性。主要缺点是消息传递能力与节点运算能力难以匹配。

(3) 机群(Cluster)。Cluster 是用高速互连网络连接起来的一组微机或工作站，各节点都是独立的(有独立完整的内存和操作系统)计算机。互连常采用商用计算机网络(ATM、FDDI、以太网等)，采用消息传递方式通信。它具有规模可扩展、性价比高等优点，缺点主要是多操作系统难于管理和维护，而且通信延迟大。

(4) 群聚集(Constellations)。Constellations 指以大型 SMP(处理器数目不少于 16 个)为节点构成的 Cluster，各节点间通过高速专用网络互连，也称为机群 SMP(Cluster-SMP 或 CSMP)。新近的巨型机多采用这种结构，如 IBM ASCI White 系统由 512 个节点机构成，每个节点含 16 个 Power3 处理器，节点内共享存储，节点间由交叉开关互连。

纵观超级计算机体系结构，一个明显趋势是越来越多的新系统都在向 Cluster 靠拢，其差别主要在于节点机、系统网络拓扑及互连技术。

3 超级计算在典型领域的应用情况

3.1 气象(包括大气与海洋模拟)

气象是超级计算机的传统应用领域。在 2001 年上半年的 TOP500 排行榜中，用于气象的共有 23 台。这些超级计算机广泛应用于数值天气预报、海洋动力学等方面的模拟。实际应用例子有：

美国加利福尼亚大学用计算机模拟厄尔尼诺现象的 10 年周期，其大气环流程序 AGCM 计算覆盖全球(分辨率：经度 2° × 纬度 2.5°，垂直方向 29 层)范围，海洋环流程序 OGCM 计算覆盖南北纬 60 度之间的全球海洋(分辨率：经度 1/4° × 纬度 1/2°，垂直方向 60 层)。两个程序非耦合地同时在 512 节点的 Cray T3D 上运行，获得 10.1 GFLOPS 的持续性能。

美国加州工学院喷气推进实验室地球与空间科学分部对一个海洋环流模型程序 OGCM 进行测试，在 Cray T3D 上，最大网格规模达 592 × 544 × 60，256 节点时最高性能达到 38.8 GFLOPS，是单处理器时候的 250 倍。

3.2 计算流体力学(CFD)

CFD 计算涉及航空航天型号设计、高速交通工具设计等，是超级计算应用比较成功的领域之一。在 2001 年上半年 TOP500 排行榜中，仅安装在航空航天部门的就有 11 台。实际的应用例子有：

美国芝加哥大学的研究人员在 ASCI Red 上，对使用自适应网格细化的高性能反应流体模拟程序 FLASH 进行测试，计算域为 12.8 × 12.8 × 256.0 cm，分辨率从 0.1cm 到 0.012 5 cm。当处理器数目从 4 096 增长到 6 420 时，对同样问题，维持了 94% 的可扩展效率(Scaling Efficiency)。在最大规模时，6 420 个处理器达到 238 GFLOPS 的持续性能，为 ASCI Red 峰值性能的 11%。

美国工程研究与开发中心使用 Cray T3E-1200 计算水波的形成以及船只穿过河流和通道时与水波之间的相互作用，网格规模达到两亿个四面体单元，方程包含 1.7 亿个未知数，1 024 处理器时达到了 115 GFLOPS 的性能。

美国 Lawrence Livermore 国家实验室使用 ASCI Blue Pacific 计算三维湍流模型中的 Richtmyer-Meshkov 不稳定问题，网格规模最高达 2 048 × 2 048 × 1 920，使用了 ASCI Blue Pacific 的 960 个节点，运行近一个星期，获得 0.6 TFLOPS 的持续性能，产生超过 3 TB 的图象数据，分布到近 300 000 个文件中。

美国 Argonne 国家实验室数学与计算机科学分部在 ASCI Red 上进行 M6 机翼三维流场模拟，求解非结构网格欧拉方程，网格规模达 280 万个顶点，从 256 处理器扩展到 2 048 处理器的过程中并行效率达 91%，2 048 处理器时性能为 156 GFLOPS，3 072 处理器时性能达 227 GFLOPS。

3.3 核模拟

使用大规模数值计算方法求解核反应过程的数学物理模型，是超级计算的重要应用领域，是大国竞争的制高点。美国 ASCI 计划的几台超级计

算机都安装或将安装于能源部的国家实验室，主要用于核武器模拟。实际应用例子有：

美国 Sandia 国家实验室使用有限元法求解 Boltzman 方程。在 ASCI Red 上，计算规模最大达 100 000 个网格元，256 台处理机时，效率高于 80%。

美国 Argonne 国家实验室物理分部在 Chiba City(512 处理器的 Linux 机群，以太网互连)上，用 Monte Carlo 方法计算一些轻核的格林函数，在 253 个处理器时，获得 38 GFLOPS 的持续性能，加速比效率为 99.3%。

3.4 生物信息学(Bioinformatics)/医学

生物信息学成为超级计算新的应用领域，如人类基因组测序过程中产生的海量数据处理就离不开超级计算机。在医学领域，也利用超级计算机来模拟人体各个器官的工作机理及人体内各种生化反应等。该领域的应用实例有：

美国印第安纳大学的 Craig A. Stewart 等实现了 fastDNAml，一个从 DNA 片断数据推断进化树的最大可能的程序。在 IBM SP 上，对 fastDNAml 进行测试，该应用能够很好地扩展到 64 个处理器。

美国杜克大学的 Pormann 等利用计算机模拟心脏中的波阵面传导，模拟过程包括三维真实几何形体以及复杂的薄膜动力学。在 IBM SP 上，当问题规模随处理机数同步扩展时，128 处理器(网格规模 930 × 930 × 1 920)效率达 91%。另外，使用机群(100M 以太网互连)计算，32 处理器时(网格规模 560 × 560 × 576)效率达 94%。

3.5 天体物理

天体物理学家越来越多地利用计算机模拟研究天体的形成、演化与相互作用，由于计算量惊人，必须使用超级计算机。该领域应用的例子有：

美国密歇根大学的研究人员采用 BATS-R-US 程序对太阳风、磁气圈以及冠状物质喷射的效应进行模拟，求解非相对论的、可压缩的等离子体方程(结合了欧拉气动方程和麦克斯韦电磁方程组)。在 Cray T3E-1200 上使用 1 490 个处理器，持续性能达 342 GFLOPS，接近线性可扩展。

美国 Sandia 国家实验室的研究人员模拟了一颗直径为 1km、质量约 10 亿吨的彗星，以 60 公里每秒的速度和 45 度角撞击地球大气时的情形。该计算包含 5 400 万个区域，在 ASCI Red 的1 500 个处理器上运行了 48 个小时。

美国 Sandia 国家实验室的研究人员进行中子星模拟，其物理模型涉及广义相对论、流体动力学，需要求解非线性、耦合的双曲椭圆方程。在 Cray T3E 上，网格规模最大达 5.32 亿个网格点，1 024 个处理器时性能达 140 GFLOPS。

日本东京大学的研究人员进行黑洞模拟。在其模型中，银河系包含 786 432 个等质量的恒星，黑洞用占整个系统质量 1% 的三个质点建模，浮点操作总数达 1.451×10^{17}。在 GRAPE-6 上，整个模拟耗时 29.88 小时，平均计算速度 1.349 TFLOPS。GRAPE-6 是专用于天体物理 N-体计算的巨型机，测试时配置有 96 个流水线处理器，理论峰值性能为 2.889 TFLOPS。

3.6 地球物理

对一些地球物理现象如地震、地球磁场等进行模拟，是超级计算机的任务之一。2001 年上半年 TOP500 排行榜中，用于地球物理的共 13 台。应用实例有：

Paul J. Tackley 等在 NASA 资助下使用谱变换方法程序 DYNAMO 计算三维地球磁场模型，在 512 节点的 Cray T3E-600 上，网格规模 1.14 亿个网格点(nr = 129，nlat = 1 150，nlong = 768)时，性能达到 150 GFLOPS。

3.7 基本理论计算

根据物理基本理论，通过超级计算机模拟的方法来研究物理现象、物质结构和相互作用，已成为继实验方法之后的又一种重要手段。应用的例子有：

美国 Lawrence Livermore 国家实验室和 IBM 等的研究人员对铁锰钴(FeMn/Co)界面上的磁结构进行第一定律旋转动力学模拟，该量子动力学模拟涉及 2 016 个原子的 super-cell 模型，揭示了界面上磁力线的方向及形状。在 IBM SP 系统上，最高达到了 2.46 TFLOPS 的性能。

美国 NASA 的一个项目组进行爱因斯坦时空方程计算。网格规模为 644 × 644 × 1 284 时，1 024 节点的 Cray T3E-1200 获得 142.2 GFLOPS 的性能，可扩展效率达 96.2%。

美国路易斯安那州立大学、日本 Yamaguchi 大学的研究人员开发出一套原子模拟程序，用于基于时空多精度算法的物质研究。他们在 IBM SP3 上进行了性能测试，使用 1 024 个处理器时，计算规模涉及 64.4 亿个原子的 MD 和 111 000 个原子的 QM 计算，并行效率超过 90%。

欧洲麦克斯-普朗克研究所开发了虚拟激光等离子体实验室(Virtual Laser Plasma Lab)代码，对相对论激光等离子体间的相互作用进行电磁直接

模拟。在 Cray T3E-600 上,784 个处理器获得超过 80 GFLOPS 的性能。

另外,在其它很多领域,如油藏模拟、材料科学、碰撞模拟等方面,也有很多使用超级计算机进行大规模计算的实例。

综观这些应用方向及应用实例,我们可以得出一些结论:

(1)国外(主要是美、日)超级计算机应用已具有相当的规模,在国防、能源、航空航天和生命科学等关键领域,都有比较成熟的应用实例。

(2)当前的应用模拟中,计算网格可高达几百万甚至超过十亿,数据量可达 TB 级,实际性能可达 TFLOPS 量级;已经能够对一些物理、化学和武器、飞行器系统进行高分辨率、高逼真度、三维、全物理、全系统的模拟。

(3)对一些计算密集型应用,如 CFD 计算、核模拟、分子动力学计算等,处理器规模达到几百到数千个以上时仍然能够看到较好的可扩展效率。而对其他应用,如生物信息学/医学、数字图象处理、数据同化等方面,实际计算时的规模可扩展能力有限,鲜有可扩展到 1 000 个处理器以上的例子。

4 体系结构与应用之间的关系浅析

4.1 应用对体系结构要求的多样性

超级计算的应用领域大致可以分为:计算密集型应用(如核模拟、CFD、气象)、数据密集型应用(如数字图书馆、数据仓库、数据融合)、通信密集型应用(如计算机协同工作、分布式作战模拟)。有的应用兼有一种或数种特性,如"数字地球[3]"就属综合型应用,仅为了实现数字美国,就需要 10^{15} FLOPS 量级的计算能力、10^{15} B 量级的存储能力、10^{15} bps 量级的宽带网络。

应用特点各异,对体系结构也就有不同的要求。如计算密集型应用,需要紧密耦合且计算能力强大的处理器集合、高速度的内部互连网络以及适应并行计算需要的各种通信、任务调度和负载平衡软件等。对数据密集型应用,需要具有海量的内外存、强大的 I/O 处理能力,以及高效的数据索引、查询等软件工具。对通信密集型应用,需要松耦合分布式结构、高可用系统,还需要好的网络安全工具等。总之,没有哪一种体系结构能够满足所有应用的要求,从而也没有哪种单一的标准能够公正评价所有体系结构的优劣[4]。

4.2 机器的可扩展性、效率

相对于计算机的峰值性能,用户更关心实际能够使用到多少处理器,处理器数目增加时计算性能是否成比例增大,即可扩展性和并行效率的问题。超级计算机的处理器数目越来越多,对并行度的需求越来越大,需要设计可扩展的并行算法与应用程序,寻求计算机体系结构与应用间的最优映射。从体系结构方面来说,需要均衡、可扩展的结构,还要高效的资源管理系统与通信库。另外,提高单个节点的速度也能减小对节点间并行性的依赖。向量巨型机因其向量处理器的计算能力强,处理器数目少,降低了对并行度的依赖,而且向量优化技术比标量优化技术成熟,因而实际计算效率较高。

4.3 易编程性是制约超级计算机应用的重要因素

尽管过去人们在为发展好的编程方式与工具方面进行了很多努力,但所取得的成果很有限。计算机体系结构不断变化,从向量机开始出现了 SMP、MPP、Cluster、CSMP 等结构,导致编程方式不断变化,出现了向量编程、数据并行(HPF)、共享存储编程(Open MP)、消息传递编程(PVM、MPI)、两级并行模式编程(如 MPI + Open MP)等,编程方式仍然难以标准化。总的说来,超级计算机的编程仍然比较困难,阻碍了各种应用向超级计算机上的转移。需要一些工具来辅助代码向并行系统的移植,无论它是以库的形式存在还是以其他软件工程工具的形式存在。自动并行(如 SUIF[5])可能是解决并行编程问题的一条途径,但在这方面无论是理论还是软件都很不成熟。

4.4 应用对存储层次的利用关系到性能的发挥

高性能处理器与存储器之间速度的差距日趋扩大,尽管存储层次的引入在一定程度上弥补了这一差距,但程序对存储层次的利用常常差强人意。由于 Cache 不命中率高,实际应用中很多高性能处理器仅能达到峰值性能的 15% ~ 25%,而应用却有使存储层次的高效利用更加困难的趋势。如在超级计算机上,人们越来越多地使用非结构网格,这意味着区域(x)的邻居不再是(x + 1)或(x - 1),必须使用数据指针来定义连接性,这使程序的数据局部性变得更差,而且由于指针结构的不规则性,进行 Cache 优化的难度增大;另外,人们对物理模型愈加精化,很多物理模型将每进行一个浮点运算就要加载一次数据,这样,程序

(下转第 107 页)

时间;(3)$\delta_i(P)$与 $\delta_i(k)$相等;(4)P 必是某个高优先级任务请求的释放时间;(5)以 P 为起点的优先级为大于等于 i 的忙周期,必包括 $S_i(k)$。因此,可以得到 P 的计算公式:

$$P = \sum_{m=0}^{k-1} C_i^{m \bmod N_i} + \delta_i(k) + I_i(P) \quad (11)$$

$$F_i(k-1) \leqslant P \leqslant S_i(k) \quad (12)$$

3.4 算法实现

算法 Compute-Rik

输入:任务集为 R,任务数为 n,任务集中任务按优先级从高到低的顺序排列。

输出:任务集 R 中任务请求的响应时间。

{for $i = 2$ to n do

计算任务 r_i 的最大响应时间

while $k < h_i/T_i$

{根据公式(8)、(9),计算出空闲时间的上下限 $\delta_i max$、$\delta_i min$。

while $\delta_i min < \delta_i max$ do

{$\delta_i = \delta_i min + (\delta_i max - \delta_i min)/2$

将 δ_i 带入公式(11)、(12)得到 P 的估算值。

判断 P 是否是某个优先级大于等于 i 的起点。

如果 P 不是某个优先级大于等于 i 的任务请求的起点,then P = 最近一个优先级大于等于 i 的任务请求的到达时间。

判断以 P 为起点的忙周期是否包含 $S_i(k)$。

if 以 P 为起点的忙周期不包含 $S_i(k)$,

then 则说明 P 取得偏小,$\delta_i min = \delta_i$ 并更新 P 值。

Else 说明 P 为最后一个忙周期的起点。

}

利用公式(4),计算出 $F(k)$的值。

$R_i(k) = F_i(k) - S_i(k)$

}

4 结束语

模拟试验表明,本文给出的计算任务请求完成时间的公式具有表达简单、计算量小的优点。对本文所给公式进行适当修改,就可扩展到通用周期多帧任务模型。

参考文献:

[1] M Joseph, P Pandya. Finding Response Time in a Real-Time Systems[J]. BCS Computer Journal, 1986,29(5):390 - 395.

[2] K W Tindell. An Extendible Approach for Analyzing Fixed Priority Hard Real-Time Tasks[J]. Real-Time Systems Journal, 1994,6(2):133 - 151.

[3] Guillem Bernat Nicolau. Specification and Analysis of Weakly Hard Real-Time Systems:[PHD Thesis][D]. Univerisitat de les llles Barlears Department de Ci[AKe·]ncies Matemàytiques I informàtica, Spain, 1998.

[4] Aloysius K Mok, Deji Chen. A Multiframe Model for Real-Time Tasks[J]. IEEE Trans on Software Engineering, 1997,23(10):635 - 645.

[5] Jose M Lopez, Daniel Garcia. A Flexible Model of Time Constraints for Control and Multimedia Real-Time Systems[A]. The 3th Int'l of Workshop on Active and Real-Time Database System[C]. Schloß Dagstuhl, Saarland, Germany, 1999.

(上接第 100 页)

的存储 - 计算比增加。除了采用更加能够容忍延迟的体系结构(如多线程)外,还需要提供低级的软件层,允许程序对 Cache 的预测和完全控制,方便用户或编译器进行存储优化。

4.5 很多应用需要可扩展的 I/O 处理能力

随着超级计算机应用日益向数据处理、事务处理领域推广,计算机处理的数据量已经达 TB 级甚至更多,这些应用对计算机的 I/O 处理能力需求越来越大。即使在传统数值计算领域,由于计算规模的扩大,数据量也成百上千倍地增加,相对于成百上千个处理器的运算能力,串行的 I/O 处理成为整个问题求解的性能瓶颈[6]。已经出现一些可扩展 I/O 和并行文件系统(如 PPFS、PANDA),但它们的可扩展性还不高,不能满足大规模并行处理和广域协作计算的要求。新的可扩展 I/O 技术与系统正在开发之中[7]。

5 结束语

随着广域网络速度的提高,超级计算将向网格(Grid)计算发展,超级计算机将成为 Grid 环境中的服务器[8,9]。在超级计算机体系结构方面,CSMP、向量巨型机都还有发展空间。此外,新的体系结构如 HTMT(Hybrid Technology Multithread,简称 HTMT)[10]、SMASH(Simple Many Self-Healing,简称 SMASH)等将得到发展。

在应用方面,超级计算应用将会开辟新的应用领域,更多地向数据处理、事务处理领域延伸,并从单纯的信息处理向知识获取发展。

参考文献:

[1] http://www.top500.org/,2001 - 12.

[2] http://www.sc2001.org/,2001 - 12.

[3] 张锁春.计算物理→科学计算→战略计算[EB/OL].http://www.bast.net.cn/kjbk/jssxtx/2000.2/b0804 - 01.htm,2001 - 12.

[4] Erich Strohmaier, Hans W Meuer. The Changing Faces of Supercomputing[J]. Scientific Computing WORLD, 2000, 10 - 11:12 - 15.

[5] http://suif.stanford.edu/,2001 - 12.

[6] Mustafa Uysal, Anurag Acharya, et al. Requirements of I/O Systems for Parallel Machines: An Application-Driven Study[R]. Technical Report CS-TR3802, University of Maryland, 1997.

[7] Tyce McLarty, Richard Hedges, et al. Integrated Computing & Communication Scalable I/O Project[EB/OL]. http://www.llnl.gov/icc/lc/sio/,2002 - 04.

[8] Ad Emmen. It's Supercomputing-It's Parallel Computing-It's Meta-computing-No, it is The Grid[EB/OL]. http://www.llnl.gov/icc/lc/siop/,2002 - 04.

[9] 徐志伟.应用为先天地宽——谈高性计算机的应用和发展趋势[EB/OL].http://www.originecom.com.cn/article/apply-first.htm,2001 - 12.

[10] Loring Craymer, Larry Bergman, Thomas Sterling. HTMT Phase 3 and Gigamesh Transition[EB/OL]. http://ess.jpl.nasa.gov/subpages/reports/01report/HTMT/HTMT - 01.htm,2002 - 02.