第 19 卷 第 3 期 郑州工业大学学报 Vol .19 No .3
1998 年 9 月 Journal of Zhengzhou University of Technology Sep . 1998

高性能并行计算机体系结构和典型的 DSM 系统—SPP 1200/XA

王文义 王若雨
(郑州工业大学计自系) (河南省电力职工大学,郑州,450001)

摘 要 详细介绍了美国在高等并行计算机系统结构领域的发展动态和 ASCI 跨世纪计划的执行情况,同时也介绍了 1 种典型的最新型分布式共享存储系统 DSM(Distributed Shared Memory)—SPP 1200/XA。

关键词 可扩展性;可编程性;分布式共享存储;ASCI 计划
中图分类号 TP338.6

在当今众多的计算机大家族中,最能代表计算技术发展水平的大概莫过于巨型计算机了。它被公认为是衡量 1 个国家的科技水平和综合国力的重要标志。过去的十年,是高性能计算机,尤其是并行计算机飞速发展并走向成熟的十年。在这期间,处理器芯片的性能翻了几番,如 HP 公司新近推出的 PA—8500 处理器,仅片上的 Cache 就有 1.5MB,而由 DEC 公司开发成功的 Alpha 21164 微处理器,则集成了 930 万个晶体管,单片性能达到了 1Gflops (每秒 10 亿次浮点运算)。计算机科学家还发明了虫蚀寻径(Mormhole)技术,找到了更符合实际的 LogP 并行计算模型。所有这些都为研制跨世纪的超级巨型计算机奠定了基础。

1 并行计算机的两大要素——可扩展性和可编程性

如果从存储结构和编程模式来划分,并行计算机可分为共享存储多机系统(Shared memory MultiProcessing ,SMP) 和大规模并行处理系统(Massively Parallel Processing , MPP) 两类。这两类计算机在性能上互有长短。

1.1 SMP 系统具备良好的可编程性,但不易扩展

SMP 系统为共享存储结构,一般比较适用的是由 4 个,8 个,16 个或 32 个处理器所组成的系统,但处理器最多不超过 100 个。程序设计仍以传统的高级语言为基础,由系统提供自动并行识别或增加并行语言成分。其特点是数据共享容易,编程方便。这种结构的缺点是系统的可扩展性差。同时,存储器访问的容许时延也是主要的限制。因此,SMP 系统难以满足当今世界级挑战性课题对高性能计算机(3T 指标) 的需求。

1.2 MPP 系统具有良好的可扩展性,但不易编程

MPP 系统一般采用分布式存储结构,可由成百上千乃至上万个处理器互连在一起组成大规模并行处理系统。MPP 的体系结构,由于在理论上对处理器个数几乎没有限制,根据

收稿日期:1997—12—10
第一作者 男 1947 年 2 生 学士学位 副教授

实际需要可随意增减,因此具有很好的可扩展性,被国际计算机界公认为是实现高性能超级计算机(3T 指标)的主要途径。

正是由于 MPP 系统的存储器是分布在多个处理器结点之上的,各结点之间需通过互联网以消息传递(Message Passing)方式交换信息,这就给程序设计带来了相当高的难度,因为编程时必需要考虑任务划分,互联网结构,cache,流水线等一系列的因素,这显然已远远超出了传统意义上的编程模式,结果是引起 MPP 系统普遍存在着用户实际有效速度与系统的理论峰值速度大相径庭的问题。鉴于这种情况,美国的一些超级计算机用户,如能源部,国防部和太空署的部分大型科研项目已开始转而使用日本的 VPP 500,从而引起了朝野上下极大地震动! 美国国会曾于 1996 年底组织了 20 几位资深专家对 MPP 研究计划的实施情况进行了调查评估,结果认为 MPP 存在:(1)太大的 Cache 延迟(Cache Latency);(2)太大的消息传递延迟(Message Passing Latency);(3)潜在的译码延迟(Decode Latency)等主要缺陷。

1.3 分布式共享存储结构(Distributed Shared Memory,DSM)

在并行机 20 年来的发展中,基本遵循的规律是,当可扩展性和可编程性发生冲突时,将首先选择可编程性而舍弃或适当兼顾可扩展性。随着科学技术的发展,应用领域对计算机性能的需求急剧增长,使得上述规律已远远不能适应。并行机到底向何处发展,能不能找到 1 种既易于扩展,又易于编程的体系结构? 答案是肯定的。著名美籍并行处理专家黄凯教授早在 1993 年就指出:"并行处理的发展趋势是用分布式共享存储结构 DSM 和标准 Unix 来构造可扩展超级计算机"。今天看来,他的预言是完全正确的。因为近年来面世的多种并行计算机,如 KSR 公司的 KSR1 就是率先采用 DSM 的 MPP 系统,此外,SGI/CRAY 公司的 S2MP(Scalable Share Memory MultiProcessing) Origin 系列和 HP/Convex 公司的 SPP(Scalable Parallel Processing) 2000 Exemplar 系列,以及下文中涉及的一些所谓跨世纪高性能计算机也都基本上采用了这种结构。

2 ASCI 计划——美国著名国家实验室和计算机工业巨头的联手跨世纪行动

ASCI 计划是在美国签署了全面禁止核试验条约的背景下启动的。为了保证核武器的绝对安全性和可靠性,就必需掌握和进行 3 维模拟计算。为此,美国能源部支持在 10 内年投资 10 亿美元,以便为虚拟核试验和为 3 维模拟计算称为"基于科学的库存管理"(Science—Based Stockpile Stewardship)问题提供足够的计算能力,预定目标是在 2002 年左右研制出每秒一百万亿次的超级计算机。这样的计算机能把非核物理和高能爆炸的实验模拟与核物理研究、高级材料、化学工程方面的技术结合起来,应用于核武器管理和科学工程计算,并能可靠地取代实际的地下核试验。

ASCI 计划分两个阶段实施:第 1 阶段目标是于 1998 年实现每秒 3 万亿次的超级计算机(表 1,表 2)。其中 Intel 公司已先于 1996 年 12 月推出了安装在桑迪亚国家实验室的名为 ASCI Red 的新型超级计算机,性能指标为每秒 1.8 万亿次,它由 9000 多个 Pentium Pro(P6)处理器组成。第 2 阶段则将于 2002 年左右实现每秒一百万亿次超级计算机的目标。

表 1　美国 ASCI 计划支持的三台超级计算机

研制公司	安装地点	拨款（千万美元）	性能指标（万亿次/秒）	处理器（个数）	完成日期	主要应用	其它应用
Intel	桑迪亚国家实验室	5.5	1.4	Pentium Pro (P6)(9000 多)	1996 年 12 月 16 日	核爆模拟	天气自然灾害预报 基因研究 太空模拟 大规模科学计算
IBM	劳伦斯利弗莫尔国家实验室	9.3	3.2	Power PC (4096)	1998 年	核爆模拟	汽车碰撞 医疗过程 飞机飞行 地震与气侯
SGI /Cray	洛斯阿拉莫斯国家实验室	11	3.2	R10000 (3072)	1998 年 12 月	核爆模拟	工业过程 全球气侯 生物技术

表 2　IBM 和SGI /Cray 公司分阶段实现计划

研制公司	计划时间（年）	性能指标（亿次/秒）	CPU（个数）	内存（GB）	硬盘（TB）
IBM	1996	1360	512	100	2.6
	1997	5650	1024	200	
	1998	32000	4096	2500	75
SGI /Cray	1996	1000	256	128	2.5
	1997	4000	768	256	2.5
	1998	32000	3072	500	75

3　一个典型的DSM 结构——SPP 1200/XA 系统

SPP 1200/XA 是由 HP 和Convex 两家公司高技术合作的结晶，是 1 种新型的高性能可扩展并行处理系统(Scalable Parallel Processing ，SPP)，采用分布式共享存储结构，超结点(Hypernode)是它的基本部件，通过一致的超环面互连(Coherent Toroidal Interconnect ，CTI)来连接各个结点，构成整个系统(图 1)。超结点中的全部CPU 、内存和I /O 由高带宽的多路交叉开关连接。由CTI 连接不同数目的超结点，组成不同可选规模的全局共享内存多处理机系统。超结点内的CPU 是紧耦合的，支持细粒度并行，超结点之间可通过共享内存和/或显式地消息传递机制实现中粒度并行。系统目前可扩至 128 个甚至 512 个CPU 。处理器采用 HP PA －7200，其峰值速度为 240Mflops 。存储器采用六层结构，分别为：(1) Cache (0)：与CPU 在同 1 块板上的存储器；(2) 线程私用(1)：专用于 1 个执行线程的存储器；(3) 结点私用(2)：专用于单一结点所有线程的存储器；(4)CTI Cache (3)：用于CTI 传输数据的存储器；(5) 近共享(4)：驻留在与进程请求所处结点相同的结点上的全局存储器；(6) 远共享(4)：驻留在与进程请求所处结点不同的结点上的全局存储器。

其中 0～3 层是局部存储器，4～5 层是全局存储器。

系统的程序开发环境包括并行化编译程序，消息传递程序库和优化的数学程序库等。此外，SPP 1200/XA 还配有优秀的可视化系统分析工具 CXPA(Convex X—window Perfor-

mance Analysis tool)，它能以图形方式向用户提供诸如程序并行度，Mflops，CPU 时间，Wall Clock 时间(即生命时间)，Cache 缺失率以及各处理器间的通信或数据延迟等信息，用户可以据此去分析和改进自己的程序性能。

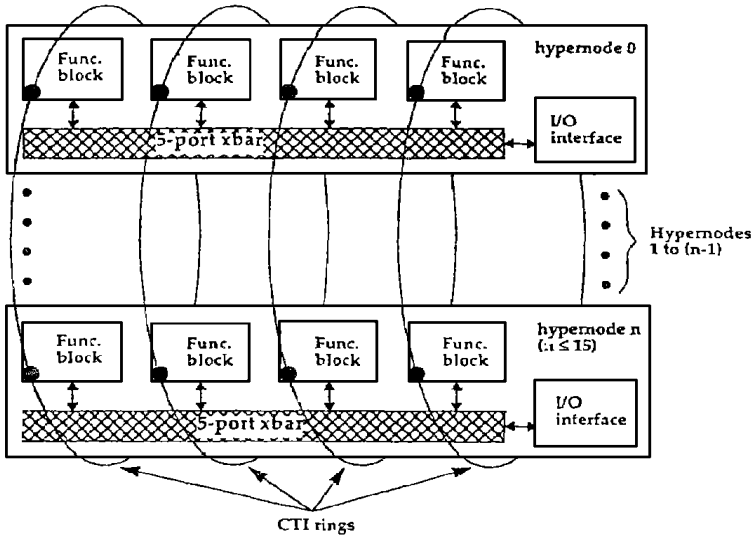

图 1 SPP－1200/XA 结构

4 结束语

因受各种因素的制约，很多人只能在微型计算机上或工作站上从事自己的工作而很少了解有关并行计算机的知识，当然也更谈不上对其系统结构有所研究，而并行计算机又的的确确与国家的发展过程息息相关。作者谨把自己掌握的有关并行计算机发展动态的最新信息整理成文，以飨读者。

参考文献

1 Kai Hwang. Advanced computer architecture :parallelism scalability programmability. McGraw－Hill ,1993: 3～25,295～301

2 Bonniger T,Esser R,Krekel D. CM－5E ,KSR2, Paragon XP/S :A comparative description of massively parallel computer. parallel computing,1995,21,199～232

3 SPP 1200/XA scalability computing system. Convex computer corporation ,Richardson ,TX,1995:34～78

4 王文义. 世界级重大挑战性课题与大规模并行处理. 郑州工业大学学报,1997(4):87～90

High Performance Parallel Computer Architecture and a Typical DSM System—SPP 1200/xa

Wang Wenyi

(*Zhengzhou University of Technology*)

Wang Ruoyu

(*Hennan University of Electric Workers*)

Abstract This paper introduces in detail the recent development of the American advanced parallel computer architecture and the implementation of the ASCI across－century project. At the same time,it also introduces a typical latest distribution share memory system—SPP 1200/XA.

Keywords scalability ;programmability ;distribute share memory ;ASCI project