超级计算机和高度并行计算技术

BIIL BUZBEE

(NCAR，电子计算机室)

提要：现在，地球科学家正试图解决一些问题，而这些问题即便利用当前最高档的超级计算机来处理，也会感到十分棘手。为了进一步提高计算机的处理速度就必须引入并行处理技术。本文引用Ware的并行处理模型旨在说明：在所用的各个模型完全是高度并发的(highly concurrent)并且通过任务管理系统减少额外开销的条件下，使用含16个处理器并共享一个存储器的超级计算机可将计算速度提高一个数量级。本文还介绍一种高度并行阵列结构的计算机——联接机(CM)，以及简述开发这种计算机的潜力以提供比当今计算机更高的计算速度的计划。

1. 引　言

长期以来，在大气科学中人们广泛应用各种高级计算机对天气、气候、海洋及有关现象进行模拟。然而，正如Holland和McWilliam[1]所说，一个模式的性能常常与所用计算机计算能力有关并受其制约。换言之，只有性能更好的计算机才能产生更逼真的模式。例如，15年前利用CDC 7600型计算机进行气候模拟，每作24小时的模拟则要1小时的CPU时间。因此，一般在多年一月份平均状况下(即用一月份通常的天文、化学和海洋强迫作用)，建立典型的模式，最多只能模拟随后几个月的情况。应用国家大气研究中心（NCAR）安装的高速向量机CRAY-1，每模拟24小时大气状况所需的CPU时间不到2分钟。这促使气候研究工作者们除了在多年一月平均情况下模拟外，敢于作其它情况的模拟，并且研究诸如大气的时间平均状态的年际变化机制这样的问题。不过，对于20年的模拟需要CPU时间长达200小时。在NCAR安装一部拥有256M字固体存储设备（SSD）的计算机CRAY X-MP/48后，人们对这台机器的应用开拓了模式模拟的科学潜力。例如，科学家现正应用X-MP型计算机来研究地球的轨道变化与18,000年前最后一次冰川期之间的潜在关系。这样的一次模拟需要在X-MP的一个处理器上运行30至40个小时的CPU时间。Raymond Roble[3]和他的合作者们使用X-MP研究热层对太阳能量输出变化的响应。这样的一次模拟需要一个多小时的CPU时间及至少1M字的内存。他们最终希望通过这些模拟试验能更好地理解高层大气的无线电信号的传播、卫星所受的阻力和化学物质输送等方面的问题。Terry Clark[4]和他的合作者应用X-MP作云模拟。他们的积云卷夹计算要用到20至30个变量，而每个变量需2M字的内存。同时，典型云的模拟过程大约需20小时的CPU时间。Robert Chervin和Albert Semtner[6]利用CRAY X-MP/48的并行处理能力分别研究他们各自设计的气

候和海洋模式。虽然这两个模式的计算都以超过 400 Mflops（每秒百万次浮点运算）的高速进行，但也需要运行 100 多个小时的墙上时钟时间。因而，自电子计算机时代开创以来，大气科学家们总是不断地渴望使用最高级的计算机，并力求最大限度地利用它们的高速计算能力。

2. 超级计算机与高度并行计算技术

并行处理器的体系结构都是多维的，主要有两种：一是共享存储器的超级计算机系统（SMSC），另一是分布式存储器的微处理器系统(DMMP)。在当前的工艺水平下，许多超级计算机采用 SMSC 结构(见表 1)，在这种系统中一些快速但昂贵的处理器共享一个大的存储器。而高度并行处理器系统则常采用 DMMP 结构，即由成千上万的相对慢些但较便宜的处理器构成，这些处理器拥有各自的存储器并通过一些通讯网络使之相互连接。目前 SMSC 和 DMMP 系统似乎可为大的模拟过程的并行处理技术的成功应用提供最大的可能性。因而，在本文下面各节，我们首先讨论由 Ware 发展的估算并行处理器运行速度的模型，然后我们将为 1990 年的超级计算机的运行速度提出预测，最后讨论高度并行系统的潜力。

表 1　当今超级计算机的性能比较

机　　型	系统周期(ns)	最大速度 (Mflops)	内存容量 (MB)	扩展存储器容量 (MB)	处理器数目
Amdahl 1400 (Fujitsu VP-400)	15(7.5)	1066	256	—	1
CRAY X-MP	8.5	235/处理器	128	4096	1,2,4
CRAY Y-MP	6	333/处理器	256	4096	8
CRAY-2	4.1	488/处理器	2000	—	4
CRAY-3	2	1000/处理器	16000	—	16
ETA-10	7~10	1250/处理器	20～18*	—	2,4,6,8
Hitachi S-810/20	14	630	256	1024	1
Hitachi S-820/80	4	3000	512	12000	1
IBM3090/600	17.2	696	128	256	6
NEC SX-2	6	1300	256	2048	1

* 每个处理器还有32M字的局部存储器。

3. 评估并行计算速度的 Ware 模型

为了评估并行计算对运算速度的提高程度，Ware[7] 开发了一个简单而有效的并行系统的速度模型。在以下的讨论中我们用 S_p 表示倍速因子，其定义为：

S_p = 最熟知的串行处理系统的运行时间/含 p 个处理器的并行处理系统的运行时间　(1)

对某个特定的计算过程，我们定义 α 为整个计算工作中可进行并行处理的部分。Ware 的模型假设在任一时刻，或者所有的处理器都在运行，或者只有其中的一个在运行，也就是说该模型假设是一个只有两种状态的机器。据此，如进行串行处理所用时间规一化为单位 1，则：

$$S_p = \frac{1}{(1-\alpha)+\frac{\alpha}{p}} \qquad (2)$$

从上式看出，在分母中的第一项是不能并行处理部分的运行时间，第二项是可并行处理部分的运行时间。S_p 是怎样随 α 而变化呢？我们在 $\alpha=1$ 处对 S_p 微商得到：

$$\left.\frac{\partial S_p}{\partial \alpha}\right|_{\alpha=1} = p^2 - p \qquad (3)$$

由此可见，在 $\alpha=1$ 附近 S_p 的改变率约为 p

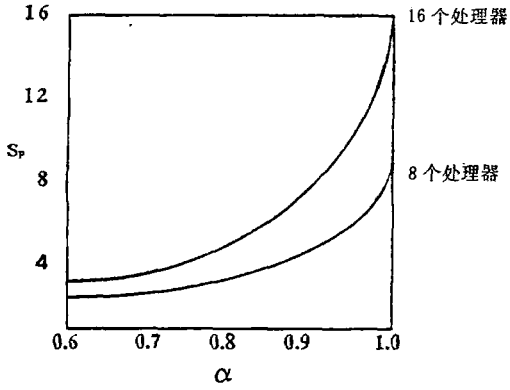

图 1 倍速因子 S_p 作为α(即可并行计算的部份)的函数的变化情况。图中给出 $p=8$ 和16的结果。

的平方。这种特性如图 1 所示。在该图中，以 α 为函数自变量并选择了两个 p 值作为比较。由以上讨论很容易得出以下结论：提高运算速度必须实现高度的并发性。因此研制高度并发的算法是成功地实现并行处理的必经之路。

4．改进的Ware模型

Ware 的模型用处很大，但它仍是不适当的。因为它假定并行处理系统与串行处理系统做同样的工作量，亦即，执行同等的指令流。而在实际工作中，这是不可能的，因为并行处理需要增加其他的指令来完成处理器之间的通讯和同步控制。而且，一个高度并发的算法可能要比最熟知的串行算法包含更多的运算过程。正因如此，我们需要在 Ware 的模型中的并行处理时间中增加第三项来改进这个模型。这一项称之为 σ，表示并行计算中的附加工作时间或额外开销。如下式所示：

$$S_p=\frac{1}{(1-\alpha)+\frac{\alpha}{p}+\sigma} \qquad (4)$$

我们注意到当 p 增加时，分母的 α/p 项将变得不重要而σ却有增加的趋势。这样，S_p 则如图 2 所示。在实际工作中也确实观测到这种速度下降的情况。如果想进一步了解并行

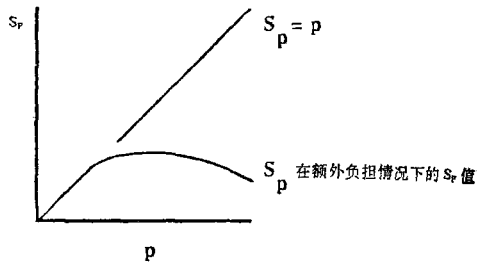

图 2 倍速因子作为处理器数目的函数。图中给出了两种情况，一种含额外开销的影响，另一种是 $S_p=p(\alpha=0,\ \sigma=0)$的情况。

计算的速度模型，可参考Flatt 和 Kennedy[8]的工作。

因此，为了在并行处理中实现高速度，采用的模式必须严格满足既增加并发性又要减少任务管理的额外开销的“边界条件”。这些要求都对问题的公式表达、算法和系统软件的设计有深远的影响。

5．共享存储器的超级计算机计算速度的发展趋势

1990年的超级计算机的计算速度将是多少？根据现有的情况，我们试图对这问题作个简要和谨慎的回答。很显然，超级计算机的运算速度与它的应用领域、实现方式和机器结构有很大的关系。因而，度量和评估超级计算机的速度是项十分复杂的工作。实际上，这是一个日益重要的研究领域。当然，超级计算机经常用峰值速度作为其重要特征，但正如 Argonne 国家实验室的 Jack Dongarra 发现的，“峰值速度就是计算机厂商所保证不会超过的速度值。”

CRAY X-MP 计算机有些硬件装置，通过这些装置可监视它在运行过程中各方面的性能。由于该机器已经运行了很长时间，因而积累了足够的有关计算机速度的数据。例如，NCAR 科学计算机研究室 (SCD) 的 Richard Sato[9] 曾对几天的时段内作过一些统计。他的统计结果表明：在NCAR的广泛应用领域中，单个X-MP处理器的计算速度平均约为 50 Mflops。这里需强调一下，这

是单个处理器在长时间间隔内的平均性能，但不能与其最优性能相混淆，因为单个X-MP处理器常常以140到200 Mflops的速度进行线性代数的软件的运算[10]。

诸如Lubeck, Moore和Mendez[11]，以及由Kampe和Nguyen[12]所作的对比研究，都表明其他超级计算机的单处理器的平均速度不会超过X-MP的两倍，因此我们将用X-MP作为估算1990年计算机速度的参照物。

有两个因素制约着单处理器速度的提高，这就是信号的传播速度和计算机的组装密度。光子的速度大约为每毫微秒1英尺，而电子的速度要慢一些。这样一个简单的事实并与表1所示的周期相结合使我们洞悉超级计算机遇到的组装上的问题。为了达到高速度，各电子元件就要用大功率驱动，而大多数元件在这种情况下都要散热。因此在超级计算机的处理器设计时要遇到如何“在用动力发动和冷却这一庞然大物的同时缩短其中最长的导线”的难题。因此，在今后5年内要想大幅度地提高单处理器的速度是十分困难的。我们保守地估计到1990年超级计算机的单处理器速度最多比现在增加一倍。我们以50 Mflops代表当今处理器速度，那么那时最多达到100 Mflops。

SMSC结构的系统现已投入使用，一些组织例如NCAR和ECMWF，正开始使用这些系统旨在并行处理大的应用项目。在NCAR，一个大型的气候模式是由三维网格组成，使用该模式进行一次典型的模拟就包含大约400万个网格点，而每个网格点要取5个变量[5]。这样把模式从某时间向前积分到下一时间就需要计算2,000万个变量。而且每当模式变量的历史文件一旦产生就有约300M字节的信息需要存储。虽然这个模式已经进行了优化，应用向量运算和并行处理，并且在NCAR的CRAY X-MP/48上以超过450 Mflops的平均速度运行，但是按科学家的要求去计算一些有代表性的问题时，则仍需墙上时钟时间100多个小时。到1990年，并行处理将可能是多处理器的超级计算机普遍采用的技术，所以我们不禁要问：这项技术能把计算机速度提高多少？

在过去的5年中，在使用有4个处理器的SMSC系统进行科学计算的过程中，已经积累了许多关于并行处理技术的数据。平均而言，使用4个处理器的系统的速度比单处理器提高了大约2.75倍。到1990年，可使用含16个处理器共享一个存储器的系统。至少有一个报告(Williams和Bobrowicz[13], 1985)指出：16个处理器的系统计算速度可达单处理器系统的10到15倍。这很容易用Ware的模型验证。例如，使用4个处理器时，倍速因子为3.75，则可推得$\alpha=0.98$。于是用16个处理器和$\alpha=0.98$，就可推出倍速因子大约是12。按改进的Ware模型，我们将假定该因子为10，并且希望这是个保守的估计。从改进的Ware模型可知，在并行处理中为了达到10倍的提高率，我们必须认真对待并行处理以及任务管理带来的额外开销。

把单个处理器的速度提高率2倍和使用并行处理技术得到的提高率10倍相结合，便可得出在今后5年中，计算机处理速度可提高大约20倍。这样的提高幅度，再加上容量惊人的存储器，将使一些我们今天的计算机仍无能为力的问题得以解决。为此只要我们愿意在SMMC系统上采用并行处理技术，那么90年代将是令地球科学界兴奋的时代。

6. 高度并行的计算机系统

高度并行的计算机系统以相对低的造价提供高的运算速度。高的运算速度可通过把成千上万的处理器组装到一个系统中来实现。而通过使用大量微处理器和存储元件可降低成本。当今可利用的功能最强的DMMP系统就是TM公司的联接计算机(CM)。我们下面以CM为例说明高度并行的计算技术的潜力。

CM计算机包括64 K(65,536)个处理器以及0.5G字节的内存。它可包含多达80G字节称为"数据储藏室"的辅助存储系统。从内存到辅存的数据传输速度高达每秒320M字节,处理器之间的通讯网络设计得很巧妙,以减少我们在改进的Ware模型中曾讨论过的额外开销。CM机是一种单指令流计算机,即它所有的处理器是同步的并且执行同样的指令流。实际上,在每个系统周期中,单独一条指令从前端机传送到每个CM处理器。若一个处理器处于活跃状态(即执行状态),那么它就用它的局部存储器中的数据执行该指令;若处理器是非活跃的(非执行状态),就不执行。活跃处理器/非活跃处理器集合可按照控制说明指令和计算结果进行动态调整。按计算机的用语而论,CM机采用的是单指令流、多数据流(SIMD)的体系结构。有趣的是,多个独立处理器的系统(MIMD)经常用编程的办法使每个处理器的局部存储器中装入同样的程序。这种方法叫单程序、多数据流(SPMD)的并行计算模式。虽然SPMD提供了比SIMD更大的灵活性,但SIMD系统更易于调试。

CM机采用的技术具有创新色彩,但又有些保守。它的处理器以16个为一组装在一个芯片上,两组芯片共享一个浮点处理器,每个处理器有8K字节的局部存储器。由于CM机采用SIMD结构,数据可进入所有的局部存储器。CM机的总功耗为28千瓦,采用气冷方式。所有65,536个处理器以阵列方式装入一个边长5英尺的立方体。这样的阵列的峰值处理速度为每秒320亿次浮点运算(精度为32位),整个阵列耗资600万美元。综上所述,CM机可提供:很高的处理速度、相对低廉的成本,大容量存储器和易编程性。但是根据我们对Ware模型讨论可知高的计算速度要求在相关联的模式中的可并发处理部份要占很大比例。

如前所述,NCAR的X-MP单处理器速度平均为50 Mflops,而采用多任务结构的模型则可使该机达到400 Mflops的速度。这些数值说明地球科学界的一些模式已经实现了高度的向量化和并行化。而且,高度的向量化的模式非常适合CM机这样DMMP结构的计算机。为了验证这个结论,我们在CM机上用一个2维的浅水模式作试验[14]。我们估计这个由65,536个处理器组成的系统的速度为1.7 Gflops。NCAR和科罗拉多大学的人士正考虑建立一个应用并行处理技术中心(CAPP),且该中心拥有一台CM机。我们的研究项目包括开发一系列的越来越复杂的三维海洋模式,并从一个简单的"块结构"模式开始进行。我们的长期目标是检验CM机的能力和适用性,看其是否能承受地球科学界设计的大计算量的模式。

(参考文献略)

谢力译自«Climate and Geo-Sciences» p. 503-511, 1989, Kluwer Academic Publishers.

吕越华校