第27卷 第7期 计 算 机 工 程 2001年7月
Vol.27 № 7 Computer Engineering July 2001

· 国家级基金项目论文 · 文章编号：1000—3428(2001)07—0046—02 文献标识码：A 中图分类号：TP312;TP338.6

一个串行Fortran程序在曙光1000并行机上的并行实现

许德政[1]，赵 林[2,3]，黄刘生[2,3]，俞国扬[1]
（1.中国科学院等离子体物理研究所，合肥230031；2.中国科学技术大学计算机科学技术系，合肥230027；
3.国家高性能计算中心，合肥230027）

摘 要：使用基于MPI并行编程方法对Fortran77串行计算程序进行并行处理，主要工作在循环级的可并行性研究和并行实现。文中给出并行代码在国家高性能计算中心（合肥）的曙光-1000机上运行的并行加速比以及相应的系统并行效率。

关键词：并行计算；消息传递；MPI编程；加速比

Parallel Implementation for a Serial Fortran Code on Dawn-1000 Parallel Computer

XU Dezheng[1], ZHAO Lin[2,3], HUANG Liusheng[2,3], YU Guoyang[1]
(1.Institute of Plasma Physics, Chinese Academy of Sciences, Hefei230031; 2.Dept. of Computer Science & Technology, University of Science & Technology of China, Hefei 230027; 3.National High Performance Computing Center at Hefei, Hefei230027)

【Abstract】 In this article, we make parallelizing processing to a Fortran77 serial routine by MPI-based parallel-programming, and put emphasis upon parallelability research and implement of cycling-layer in source programme The parallelized code was debugged and executed on Dawn-1000 computer in National High Performance Computing Center at Hefei. The actual paralleling speed ratio and the relative paralleling efficiency are given.

【Key words】 Parallel computing; Message passing; MPI programming; Speedup ratio

随着社会的发展和科学技术的进步，人们对计算机的速度要求越来越高。由于单个处理机的速度提高总是有限的，通过增加处理器的数目来加快运算速度已势在必然。目前超级计算机（即大型并行机）的出现以及相关并行计算技术的研究开发，使得求解大规模、高精度的非线性数学物理方程成为现实。例如，用于研究托卡马克等离子体磁流体不稳定性问题的数值模拟[1,2]，就是一个复杂的非线性偏微分方程的求解问题。该数值模拟是了解磁约束核聚变等离子体行为的重要手段，具有计算量大，精度要求高，耗时多等特点。因此，开发与此相关的并行计算程序显得十分必要。

本文对求解托卡马克等离子体磁流体平衡与垂直位移不稳定性的（串行）计算程序作了初步的并行化。首先从剖析原串行Fortran程序入手，弄清其逻辑关系和结构特征。程序特点显现为子例程和函数调用频繁、调用层次深。程序内循环模块多，而且存在循环嵌套情形。因为并行是在给定的串行程序基础上进行的，考虑到数据流的相关性大，故不对原来的程序逻辑结构作大的改动，并行重点是放在循环块的并行化可行性分析和改写上。当遇到嵌套循环情况时，原则上只对其外层循环进行并行处理。

在基于消息传递接口（Message Passing Interface，简称MPI）编程模型[3,4]中，计算是由一个或多个彼此通过调用库函数进行消息发送、接收通信的操作所组成。本文并行代码就是使用MPI并行编程方法写成的，属于一种粗粒度型SPMD（单程序多数据流）程序。并在中国科大曙光1000并行机上进行了调试和运行，最后对实际测算的加速比与其相应的理论估算值作了比较。

1 MPI并行编程方法简介

所谓基于消息传递的并行编程[3,4]，是指用户必须显式地通过发送和接收消息来实现处理器之间的数据交换。它是大规模并行处理机MPP（Massively Parallel Processing）和工作站机群COW（Cluster Of Workstation）采用的主要编程方式。在这种并行编程中，每个进程均有自己独立的地址空间，一个进程不能直接访问其它进程中的数据，这种远程访问必须通过消息传递来实现。因为消息传递的开销比较大，所以它主要开发大粒度和粗粒度的并行性软件。

MPI是一个消息传递接口函数库的标准定义[5]。是由MPI论坛开发研制的。MPI函数库本身与语言无关，并提供了与C/C++和Fortran语言的绑定。这个定义不包含任何专用于某个特别的制造商、操作系统或硬件的具体特性。1997年修订的标准MPI-2已经提供了200多个函数。其中，有6个最基本的函数，它们是启动和结束MPI环境，识别进程以及发送与接收消息等，列举如下：

MPI_INIT ---------- 启动/初始化MPI环境
MPI_FINALIZE ---------- 结束/终止MPI环境
MPI_COMM_SIZE ---------- 确定进程数
MPI_COMM_RANK---------- 确定自己的进程表识符
MPI_SEND ---------- 发送一条消息
MPI_RECV ---------- 接收一条消息

基金项目：国家高性能计算基金资助项目(99207)
作者简介：许德政(1949～)，男，高级工程师，主要从事受控热核聚变研究工作；赵 林，教师；黄刘生，教授、副主任；俞国扬，研究员
收稿日期：2000-12-06

这6个最基本的函数形成了在MPI中编写完整消息传递程序的一个最小集合。本例在并行编程过程中，除了使用上述几个最基本函数外，还频繁地用到群集通信（Collective Communication）中的一些函数，诸如收集（MPI_GATHER）、广播（MPI_BCAST）和路障（MPI_BARRIER）等。

2 串行程序dvdix.f的结构剖析

本文作并行处理的对象是一个含有5000条语句（行）的中型Fortran串行程序。这是求解托卡马克上等离子体垂直位移不稳定性的程序[1,2]。

其核心部份是数值求解一组弱非线性常微分方程，方程的个数可以从几十个到上百个。程序中频繁地涉及子例程和函数的调用操作。"相对主从"例程或函数调用共有4个层级，整个程序的层级结构如图1所示。并行化的重点放在求解常微分方程部分的几个子例程的循环模块上。首先分析循环体内参数之间的数据相关性[4,5]。当循环体内部，循环体之间或循环嵌套之间存在交叉相关时，不作（也不能）并行改写。并行处理只对循环体间不存在相关性的情况进行。遇到循环嵌套时，一般只考虑外层循环的并行化。

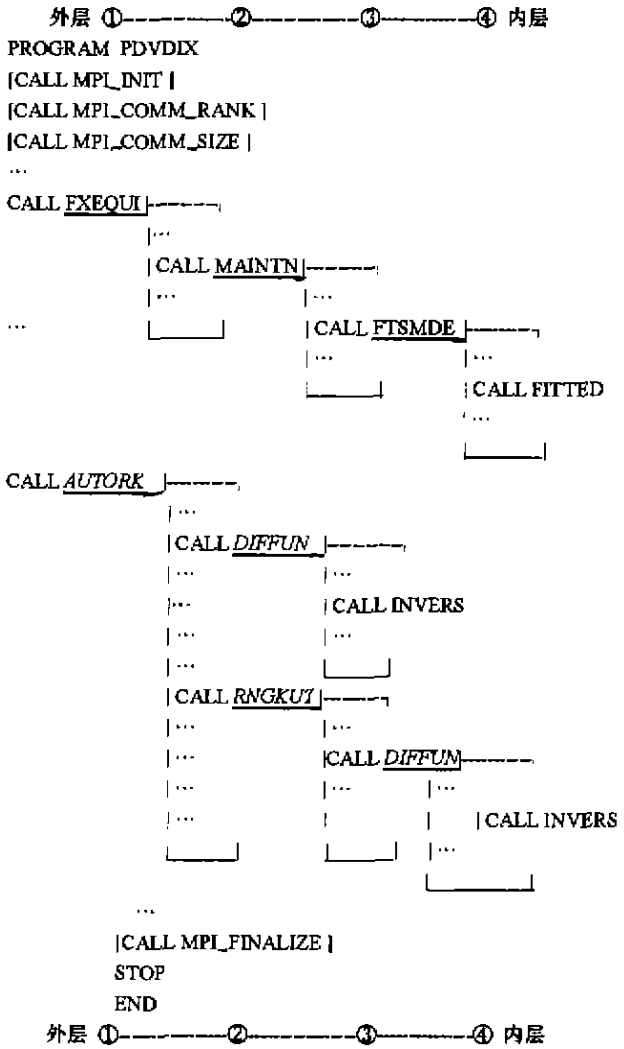

图1 串行程序dvdix.f的层级结构

图1说明：1)方括弧内的命令为并行后才添加的MPI函数调用。2)凡以叙体标示的Fortran子例程(即*AUTORK*、*DIFFUN*和*RNGKUT*)均被作了并行处理；3)省略号处代表未标出的其它语句行及函数或子例程调用。

3 在曙光1000上的并行实现

曙光1000并行机[6]是国产的基于消息传递（Message Passing）、分布存储的松散耦合大规模并行处理（Massively Parallel Processing）系统。该系统共有36个节点（Node）机，其中包括32个基于i860的计算节点（Computing Nodes），2个系统服务节点子系统（又称为主机，Hosts）和由2个I/O节点组成的外存子系统（I/O Nodes）。其软件系统是基于Unix的分布式操作系统。

将并行代码放在曙光1000上进行运算。分析并行计算的结果，并与串行结果进行比较，得到并行程序的加速比S（即S = 串行执行时间 / 并行执行时间），并且评估并行系统的效率E（E = 加速比 / 处理器[节点]数）。结果显示，并行加速性能达到Amdahl定律所预期的指标。

4 并行计算结果与性能评估

Amdahl加速定律：Amdahl加速比定律[3,4]的基本出发点是：①对于很多科学计算，实时性要求很高，即在此类应用中时间是个关键因素，而计算负载是固定不变的。因此在给定的计算负载前提下，为达到计算结果实时性可通过增加处理器的数目来提高计算速度。②因为固定的计算负载可分配到多个处理器上，这样增加了处理器的个数就意味着加快了执行速度。关于固定计算负载的加速公式表示如下：

$$S = [W_s + W_p]/[W_s + W_p/p] \quad (1)$$

式中,S为加速比,p代表并行系统中的处理器个数。W_s、W_p分别表示应用程序中的串行分量和可并行化部分。

引入参数$f = W_s/[W_s+W_p]$，即f代表串行分量比例；显而易见$r = 1-f$表示并行分量比例。将$1/f = [W_s+W_p]/W_s$代入(1)式，归一化得：

$$S = p/[1 + f\ (p - 1)\ ] \quad (2)$$

当$P \to \infty$时，上式极限为：

$$S = 1/f \quad (3)$$

这就是著名的Amdahl加速（比）定律，它表示随着处理器个数的无限增大，并行系统所能达到的加速比之上限为1/f。亦即加速比S与应用程序中的串行分量比例f成反比。

本例加速效果：由于被并行处理的应用程序的计算负载是固定的，所以在曙光1000上，为加快计算速度，我们依次取处理器数目为1、4、8、16等4个不同情况分别进行计算，表1给出并行计算的加速性能结果。

表1 并行计算的加速性能

处理器数（p）		1	4	8	16
并行分量（r）		0.3			
实测值	加速比(S)	1.00	1.27	1.35	1.37
	并行效率（E）		~0.3	~0.2	~0.1
理论值	加速比（s）	1.00	1.29	1.36	1.39
	并行效率（e）		~0.3	~0.2	~0.1

表中加速比的实测值$S = T_{SERIAL} / T_{PARALLEL}$，其理论值(s)则根据上述（2）式计算。并行加速效果是明显的，而且实测值与理论值相吻合。

5 结论与进一步的工作

本文工作是在给定的串行程序基础上进行的，由于数据

（下转第87页）

器)，帧的最小发送时间2τ =51.2μ s，即最小帧长为2τ * 10Mbps=512bit=64byte。

所以对10Base-T，有表1所示的效率。

表1 以太网帧长与效率的关系

帧 长	64byte	128byte	256byte	512byte	1024byte
效率η	0.2~0.3	0.4~0.5	>0.6	0.7~0.8	0.85

按上表，当帧长为1024byte时，10Base-T的带宽为8.5Mbps，如果由10个站点共享，则平均每个站点的带宽为850kbps。

对某一特定站点来说，其等待大于m个时隙的概率

$$P=\sum_{j=m}^{\infty}\frac{A}{k}(1-\frac{A}{k})^{j-1}=(1-\frac{A}{k})^{m-1}$$

取k=10,A=1/e，则m=10时，P<0.72; m=100时，P<0.025; m=200时，P<5.76*10⁻⁴; m=500时，P<7.5*10⁻⁹；即等待超过5.12ms的概率小于2.5%，等待时间超过25.6ms的概率小于10⁻⁸，只要设计系统时考虑相应的抖动时间，就可以满足要求。

2 协议的改进

以太网上的传输层主要协议是TCP和UDP，TCP是面向连接的传输协议，提供可靠的传输，但可能时延及其抖动较大；UDP是无连接的传输协议，消耗资源少，但不可靠。这两个传输协议都不是很适合传输连续流，TCP的缺点是抖动较大，如果直接用UDP，虽然LAN上可靠性很好，误码率达10^{-6}，仍不能满足有些要求。如果把TCP和UDP结合起来，形成一个新协议，原理如下：首先采用TCP，对发送的数据进行确认，如果发送失败，则重发，但是当超过时限，则放弃。这样既可以大大降低误码率，又可以满足实时需要。以最多试发两次为例，则误码率降为10^{-12}，这可以与光纤的误码率媲美，足以满足绝大多数应用。

3 交换机性能分析

全双工交换机的广泛使用，大大提高了以太网的带宽，由于交换机不是共享型，克服了传统Hub的发送冲突的缺点。而发送冲突正是以太网不可靠、无预测性的原因。全双工交换机的带宽是Hub的K/2倍，其中K为端口数。

既然以太网交换机和ATM交换机都采用交换原理，那么为什么以太网交换机可预测性差呢？原因是ATM采用固定小数据包，而以太网的数据包是变长的，这样交换机就很难精确分配带宽。

如果我们按照以太网交换机的原理，设计数据包的大小，就可以控制带宽的分配。因为以太网交换机内部采用轮询原则，这样，只要我们设计数据包大小的平均值较大，就可以得到较大的带宽。这里需注意的是交换机的数据包小于1.5kbytes,超过了就会丢弃。采用TCP协议时下层把应用层的数据划分为适当大小的包，但UDP协议是直接发送，超过部分就被丢弃。但是也不是在1.5kbytes内越大越好，因为这和数据源有关，太大了组包时延就越大。所以要根据具体情况，权衡组包时延和带宽，然后选出一个合适的包长。

4 结束语

本文证明了以太网用于多媒体网络是可行的，作者在10Base-T LAN上进行了音频、视频的传输，对64kbps的音频，声音清晰连续，基本上和打电话没有差异。对1.5Mbps的MPEG I电影，画面平滑，无抖动。国际上，网络会议电视的ITUT-H.323标准也表示以太网传输连续流的可行性。

参考文献

1 ITU-T H.323 Visual Telephone Systems and Equipment for Local Area Networks Which Provide a Non-guaranteed Quality of Service
2 Tannenbaum A.S. Computer Networks(Third Edition) Prentice Hall, 1999-06
3 周明天 TCP/IP网络原理与技术 北京：清华大学出版社，1998
4 Newman P. ATM Lcoal Area Networks.IEEE Commnu. Magazine, 1994,32

※※※

(上接第47页)

流的相关性大，故不对原来的程序逻辑结构作大的改动，并行重点是放在循环语句的并行化可行性分析和改写上。当遇到嵌套循环情况时，只对其外层循环进行并行处理。

在向主机申请计算节点时，尽量做到节点序号的连续性和节点分区的矩形特征。本文实际取用的节点数目分别为2^k（k=0，1，2，3，4）。而且分区的选择决定于循环变量的取值。这是因为在消息传递中应考虑到数组元素分配的均匀性。

在考虑并行化的颗粒度（Granularity）时，必须在计算（负载）与通信（数据）之间寻求折中方案。因为并行度（Degree Of Parallel）和颗粒度通常是互相制约的[4]。本文的实际并行效率之所以较低，是由于整个程序的并行度不够大（并行比例系数r仅为 0.3）。

充分使用MPI群集通信子例程的强大功能，尽可能减少消息传递指令出现频度。同时要慎重使用路障（MPI_BARRIER）和等待(MPI_WAIT)函数，以避免不必要的同步开销。另外当循环体的（计算）负载远小于因并行带来的（数据）通信开销时，并行意义不大。

由于托卡马克等离子体磁流体不稳定性的数值模拟在磁约束核聚变研究中有重要意义，随着国内诸如曙光、神威、银河等高速并行计算机投入运行，我们将在目前工作的基础上，继续加大程序的并行比例，优化程序结构，以达到进一步提高系统的并行效率的目的。

致谢: 在并行程序调试和上机过程中，本文作者之一（许德政）曾与国家高性能计算中心（合肥）计永昶博士和陈志辉博士作过有益的讨论，在此表示感谢。

参考文献

1 Du T,Yu G Y Vertical Displacement Instability of Elongated Plasma and Its Dynamic Stabilization.Scientia Sinica,1992(A35) 292
2 Yu G Y.Parametrization Study of the Decay Index of the Equilibrum Magnetic Field in Tokamak.Nuclear Fusion,1994,34 1408
3 陈国良 编著.并行计算-结构 · 算法 · 编程.北京：高等教育出版社. 1999-10
4 黄 铠,徐志伟 著,陆鑫达译.可扩展并行计算-技术、结构与编程. 北京：机械工业出版社，2000-05
5 谭浩强，田淑清 编著.Fortran 77 结构化程序设计 北京：高等教育出版社.1985-04
6 国家高性能计算中心（合肥）编.曙光1000系统环境及使用指南. 中国科学技术大学.1998-12