文章编号: 1001-2486 (2000) 05-0064-04

差分离散方法分布式并行计算的重叠边界优化*

张理论[1], 叶红[2], 孙安香[1], 李晓梅[3]

(1. 国防科技术大学计算机学院, 湖南 长沙 410073; 2. 北京工商大学信息工程学院; 3. 装备技术指挥学院)

摘 要: 针对分布式存储环境下显式差分方法的并行计算问题, 依据分布式存储多处理机的体系结构特点, 提出了一个重叠边界优化模型; 该模型目前已成功应用于中国科学院第三代海洋环流模式的分布式并行计算优化。

关键词: 分布式存储; 并行计算; 差分离散; 重叠边界; 优化
中图分类号: TP301.6 文献标识码: A

Optimal Overlapping Boundaries Model for Distributed Memory Parallel Computing of Difference Discrete

ZHANG Li-lun[1], YE Hong[2], SUN An-xiang[1], LI Xiao-mei[3]

(1. College of Computer, National Univ. of Defense Technology, Changsha 410073, China;
2. School of Information Engineering, University of Business Administration; 3. Institute of Equipment Technology and Command)

Abstract: An optimal overlapping boundaries model is presented, aiming at the explicit difference schemes on distributed memory parallel computing, in accordance with the architecture of distributed memory multiprocessors. This model has been successfully utilized in the optimization of distributed parallel computing of the third generation ocean general circulation model of the Chinese Academy of Science.

Key words: distributed memory; parallel computing; difference discrete; overlapping boundaries; optimization

差分离散在偏微分方程数值求解中是应用最广泛的经典方法[1], 其中显式差分离散是一个重要组成部分。对显格式的分布式并行计算优化, 目前还鲜有相应的研究。本文讨论了该问题。在分布式存储环境下设计显式差分离散的并行算法时, 在新的时间积分开始前应更新边界数据。对大规模并行处理机 (Massive Parallel Processor) 和群机并行系统 (Cluster of workstations) 来说, 更新边界数据是通过特定节点间的显式消息传递实现的。现有的MPP和COW虽具有较高的带宽和较快的传输速度, 但消息传递开销相对于浮点计算来说仍然很大, 甚至有时会成为提高并行效率的最主要障碍[2]。分布式并行的通信开销主要来自消息投入网络和从网络接收的系统开销, 这种系统开销在消息长度不超过一定阈值 (若干Kb) 时, 几乎不受消息规模的影响[3]。考虑到边界数据交换只涉及到小消息传递 (1Kb以下), 因此其通信时间主要来自于系统开销。据此本文提出了一个显式差分方法分布式并行计算的重叠边界优化方法, 并给出了计算模型; 旨在有效减少消息传递频率, 在消息长度和传输频率以及CPU速度之间寻求一个最佳的结合。本文的优化模型揭示了该结合点依赖于并行机体系结构和应用问题的自身特点。

1 MPP/COW 的通信特点

MPP 和COW 是分布式并行计算的主要硬件平台。MPP的每个计算节点采用一个或多个通用微处理器, 节点间通过高带宽、低延迟网络连接, 并通过处理器的存储总线及网络接口线路 (Network Interface Circuitry) 进行耦合。COW 的每个节点往往是一台独立微机或工作站, 有时甚至可以是对称多处理机 (SMP), 节点连接用较经济的网络, 如以太网 (Ethernet)、光纤网 (FDDI)、ATM 交换网等; 其网络接口与节点的 I/O 总线耦合。采用 MPP 和 COW 进行并行计算的共同之处在于: 物理上的分布

* 收稿日期: 2000-04-07
基金项目: 国家 863 计划项目资助 (863-306-ZD11-0318)
作者简介: 张理论 (1975-), 男, 博士生。

存储；借助消息传递实现节点间通信。无论MPP还是COW，其通信开销相对于基本计算时间来说都是非常大的。鉴于IBM SP2兼具MPP和COW的特点，这里以它为例进行说明：SP2使用66MHz、266Mflop/s的POWER处理器，其上的一个消息通信的最短时间为39μs，这相当于2601个CPU时钟周期或是10374个浮点计算。由此，节点通信在分布式并行计算中至关重要，是降低通信时间、提高并行效率的重要途径之一。事实上，消息通信时间与消息规模并不呈严格正比例关系，这一点可以用Hockney提出的一个点点通信时间模型[4]进行说明$(t(m)=t_0+m/r_\infty$，其中$t(m)$为通信时间开销；t_0为消息发送或接受的启动开销，时间单位为μs；m为消息规模，单位为byte；r_∞为通信的渐近带宽，单位为MB/s)。现有的测试结果表明，该时间模型与实际情况比较吻合[5]。当消息长度较小(若干Kb以下)时，通信时间基本上维持同等规模。

2 显式差分的并行计算

用下式表示与时间相关的偏微分方程组

$$Lu=0$$

若问题确定且有适当的初边值条件，则可用某种显格式差分方程组来离散求解，得到

$$L_h u_h=0$$

下标h表示空间和时间的离散化参数。为阐明原理，仅以简单的波动方程为例：

$$u_t=au_x=0$$

对其做如下差分离散

$$u_t\sim\frac{u_j^{n+1}-u_j^n}{\Delta t}\qquad u_x\sim\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x}$$

则得差分方程为

$$u_j^{n+1}=u_j^n-\frac{\lambda}{2}(u_{j+1}^n-u_{j-1}^n)$$

其中上下标分别表示时间层和空间层，$\lambda=a\Delta t/\Delta x$。

对以上方程的分布式并行计算，往往进行空间分解。即将网格进行剖分后，分布在不同的处理器上。若处理器逻辑划分与网格分割同构，则相邻节点应有一条重叠网格边界。定义严格边界与伪边界：前者为各个处理器直接由数据剖分得到的网格边界；后者表示为将并行计算正确进行下去，各个

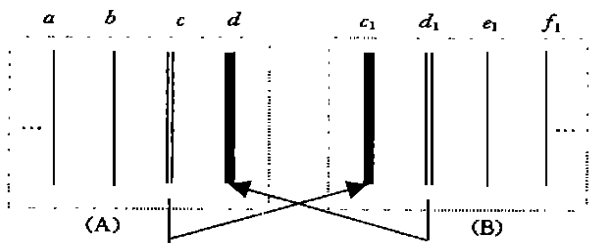

图1 伪边界交换示意图

Fig. 1 Exchange of fake boundaries

处理节点所必须延伸到的最近边界。如图1所示，在上图中A、B表示网格子区域。c为A的严格边界，d是A的伪边界；d_1是B的严格边界，c_1为其伪边界。一般地，伪边界的条数由微分方程以及采用的差分格式决定。在应用问题中，往往需要频繁的时间积分，每步一般要包括计算和通信两大部分：

- A区域计算自左至c边，B区域计算自右至d_1边；
- A、B间通信：c边传递给c_1，d_1边传递给d边；

• Barrier 同步。

上例中，单个边界重叠就能完成计算，但需要在每个计算步后进行通信。对于一些重要的应用，若单步计算时间相对较短，频繁交换边界数据就会成为提高并行计算效率的瓶颈。采用多个重叠边界是缓解瓶颈的一个自然的做法，其基本思想是在原有伪边界的基础上，增加若干条重叠边界，利用冗

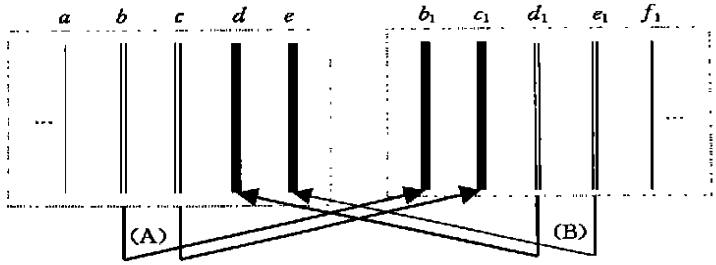

图 2 多边界交换示意图

Fig. 2 Exchange of multi-boundaries

余计算掩盖部分通信。图 2 给出了增加一条重叠边界时的通信特点，其中 c、d_1 为严格边界。若初始数据分布充分，记 STEP 为积分步数，则单步计算如下

STEP 为奇数，则 {

• 在 A 区域自左计算至 c 边，在 B 区域自右计算至 d_1 边

• A、B 间通信：b、c 边分别传递给 b_1、c_1；d_1、e_1 边分别传递给 d、e 边

• Barrier 同步 }

STEP 为偶数，则 {

• 在 A 区域自左计算至 d 边，在 B 区域自右计算至 c_1 边

}

3 优化模型

对特定的应用问题，重叠多少条边最佳？这不仅取决于问题本身的计算特性，还有赖于所采用的并行软件平台和硬件体系结构。针对以上问题我们提出了多重叠边界优化模型，依据该模型能在通信与计算之间寻求一个良好的组合，以减少总的通信开销。主要的模型参数有：N 为在伪边界的基础上增加的重叠边界层数；m 为单层通信规模 (Byte)；t_k^c 为 N 取 k 时所增加的冗余计算时间，t_k^c 依赖于 m、k、计算机速度以及问题本身；t_b 为一次同步开销，我们考虑理想的并行计算的负载分布。即在问题和平台确定的情况下，t_b 是恒定的。记 T_0 为只有伪边界重叠时的单步平均计算时间，T_k 为 $N=k$ 时的单步平均计算时间，S 为时间积分步数，不妨设 S 是 $k+1$ 的整数倍。为简化起见，假定只有一条伪边界，则

$$T_0 = S \cdot (T_{\text{const}} + t(m) + t_b)/S$$

$$T_k = \frac{S}{k+1} \cdot \left\{ \sum_{i=0}^{k} (T_{\text{const}} + t_i^c) + t[(k+1)m] + t_b \right\} / S$$

其中，T_{const} 为由问题固有的计算量所决定的计算时间，$t_0^c = 0$。目标函数 $N(t_1^c, m, t_0, r_\infty, t_b)$ 满足：

$$T_N = \min(T_i) \quad (i = 0, 1, 2, \ldots)$$

依据第 1 节中的 Hockney 模型，$t[(k+1)m] = t_0 + (k+1) \cdot m/r_\infty$

这里以 T_0 和 T_1 的比较来说明该模型的适用性，

$$T_0 = T_{\text{const}} + m/r_\infty + t_0 + t_b$$

于是有

$$T_1 = T_{\text{const}} + m/r_\infty + \frac{t_1^c + t_0 + t_b}{2}$$

$$T_1 < T_0 \Leftrightarrow t_1^c < t_0 + t_b$$

上式表明：当单边冗余计算时间小于最短消息通信时间与同步开销之和时，多重叠一条边界比只有伪边界重叠节省计算时间。需要指出，单边冗余计算时间由问题特点以及计算机速度决定，它直接依赖于通信边的规模，也即消息规模 m。单边冗余时间越短，消息长度相对就越小。

4 数值实验

采用国产四节点 MPP 并行系统和四节点 PC 机群以及 MPI 软件平台，以海洋环流模式中重力外波的四机并行计算为算例，实验结果如表 1、表 2 所示。

表 1 MPP 并行计算（单位：s）

Tab. 1 Parallel computing on MPP (unit: s)

积分步数	360	720	1440	2880
串行时间	4.7880	9.4385	19.3193	37.6250
单边重叠	1.3703	2.7476	5.6484	11.5117
两边重叠	1.3593	2.7015	5.5323	10.9456
三边重叠	1.5164	2.9766	6.1204	12.6723

表 2 PC 机群并行计算（单位：s）

Tab. 2 Parallel computing on PC cluster (unit: s)

积分步数	360	720	1440	2880
串行时间	13.2835	26.6324	53.7649	108.6537
单边重叠	6.4568	12.9154	25.8452	51.7872
两边重叠	5.7424	11.4762	22.9531	45.9088
三边重叠	5.3091	10.7133	21.7529	43.4762
四边重叠	5.8132	11.7001	23.5157	47.1094

数值计算采用蛙跳差分格式，空间离散在分布式并行计算时要求子区域间必须有一条伪边界。考虑到 COW 的最短消息通信时间与同步开销均相对地大于 MPP 的相关参数，而前者的 CPU 速度往往是后者的 2 倍以上，故多重叠边界优化模型在 COW 的应用潜力要比 MPP 大。以上两表显示了多重叠边界模型的有效性，实验结果与理论分析吻合得非常好。

参考文献:

[1] 郭本瑜. 偏微分方程的差分方法 [M]. 北京：科学出版社，1988.

[2] 张理论，宋君强，孙安香，李晓梅. 一个基于S2MP 结构的高效并行海洋环流模式 [C]. 并行计算学术会议论文集，2000.

[3] 莫则尧，李晓梅. 工作站网络环境下的并行计算 [J]. 计算机学报，1997，20 (6)：510~517.

[4] R. W. Hockney. A Framework of Benchmark Performance Analysis [J]. Advances in Parallel Computing, Elsevier Science Pub., 1993, 8: 65-76.

[5] Huang. Kai, Xu. ZhiWei. Scalable Parallel Computing [M]. China Machine Press, 1999.