大规模数据并行问题的可扩展性分析*

金之雁
(中国气象科学研究院 , 北京 100081)
王鼎兴
(清华大学计算机系 , 北京 100084)

摘　要

大规模数据并行处理的性能受到处理机数量、I/ O 速度、通信速度等多方面因素的制约。增加处理机数量或提高处理机的计算速度,可以提高计算机的整体处理速度,但是通信和 I/ O 会成为影响并行效率的主要因素。为了综合分析这些因素对计算性能的影响,用一种比较典型的大规模数据并行的计算模型,具体分析了处理机数量、处理机速度与处理机间的通信延迟、通信速率以及输入输出速度之间的关系。得到了大规模并行机的通信和 I/ O 性能与处理机速度与数量之间存在的关系。指出,增加处理机数量、提高单节点处理速度的同时,必须按照一定的关系相应增加节点间的通信性能和 I/ O 性能。单纯以增加处理机数量、提高单处理机速度提高计算机峰值速度的方法会降低系统的计算效率,不能达到计算速度与计算机处理能力同步增长的目的。

关键词 : 并行计算　可扩展性　数据并行

引　言

并行处理可以分为功能并行和数据并行两种方式,大规模数据并行是大规模科学计算的一种最常见并行处理方式。例如在求解流体力学方程时,往往将求解空间数据分割成不同的区域,用多个处理机进行处理实现并行计算。并行处理的目的在于加快处理速度,以往的性能分析侧重于对加速比的讨论[1],指出增加课题计算规模可以提高加速比。在满足时间要求的前提下,对于某些用户,对计算机进行升级的主要目的是减少计算时间;但对于很多用户,在计算时间能够满足要求的前提下,扩大计算机规模是为了提高计算精度,他们需要在规定的时间内得到精度尽可能高的计算结果。这部分用户的课题规模随并行机规模而增大,并行效率成为用户关心的主要问题,他们希望在扩大计算机规模的同时,大课题运行时间和并行效率基本不变。Gustafson 分析了计算机规模与求解问题规模同时扩大 n 倍时的加速比情况,提出了固定时间加速比的概念,即 Gustafson 定律[2]。但是它没有包括对计算机通信性能、I/ O 性能的分析。

本文以流体力学初值问题为背景,建立了一个大规模数据并行的性能分析模型,提出

* 本文得到国家"973"项目 G1998040911 课题和国家自然科学基金 69933020 项目的资助。
2001-09-27 收到,2002-05-08 收到修改稿。

时间不变,并行效率不变,并进一步假定计算、延迟、通信、I/O 时间都不变,在此条件下,分析了并行计算机处理机速度、处理机数量和通信延迟、通信速率以及输入输出速率之间的关系。

1 数据并行计算模型

初值问题是已知某空间在初值时刻的状态,求解经过时间 T 以后的状态。假定求解空间为边长为 L 的正立方体(图 1),用间距为 d 的三维网格进行离散化,构成格点数为 $(L/d)^3$ 的三维网格。处理机数量为 $P = n \times n$,对区域进行二维水平分割,每个区域是格点数量为 $\frac{L^3}{n^2 d^3}$ 的柱形,在水平边缘有宽度为 e 个格距的重叠区域,在进行计算时需要用到该区域的数据,在区域中心有 $\frac{L}{d}(\frac{L}{dn} - e)^2$ 区域,它计算的数据已经在该区域内。其时间分辨率为 Δt,经过 $T/\Delta t$ 步计算,得到 T 时刻的状态进行输出。

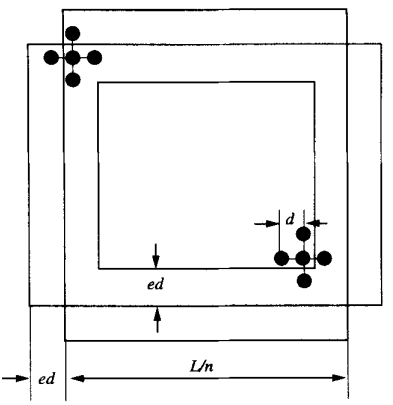

图 1 子区域结构

设每个网格点的计算时间相同,为 t_{ca}。每次通信的延迟时间为 t_{la},每个格点数据的输入输出时间为 t_{io},每个格点数据的通信时间为 t_{co},各通信链路之间无冲突,则计算时间为 $T_{ca} = \frac{T}{\Delta t}\frac{L}{d}(\frac{L}{dn})^2 t_{ca}$,延迟时间为 $T_{la} = \frac{4T}{\Delta t} t_{la}$,通信时间为 $T_{co} = \frac{T}{\Delta t}\frac{L}{d}\frac{4eL}{dn} t_{co}$,输入输出时间为 $T_{io} = (\frac{L}{d})^3 t_{io}$,时间分辨率 Δt 由计算稳定性确定,正比于空间分辨率,设 $\Delta t = \alpha d$,e 由采用的计算方法确定。总时间为:

$$T_{otal} = T_{ca} + T_{la} + T_{co} + T_{io} = \frac{TL^3}{\alpha d^4 n^2} t_{ca} + \frac{4T}{\alpha d} t_{la} + \frac{4eTL^2}{\alpha n d^3} t_{co} + \frac{L^3}{d^3} t_{io} \qquad (1)$$

2 可扩展性与运行时间的关系

Nussbaum 和 Agarwal 提出了并行计算机的可扩展性定义[3],在理想情况下,可扩展性与并行效率的定义是一致的[4]。本文建立的计算模型属于理想情况,因此,在以下分析中,可扩展性与并行效率相同。

定义 1:对于给定算法 κ,P-处理机系统的加速比 $S_{peedup}(\kappa, P)$ 和效率 $E_{ffecient}(\kappa, P)$ 分别为:

$$S_{peedup}(\kappa, P) = \frac{T_{ime}(\kappa, 1)}{T_{ime}(\kappa, P)} \qquad E_{ffecient}(\kappa, P) = \frac{S_{peedup}(\kappa, P)}{P}$$

其中,$T_{ime}(\kappa, P)$ 是在 P-处理机系统的并行执行时间。

算法规模可以定义为该算法的计算量 ,在单处理机条件下 ,算法规模增加 m 倍 ,记为 m^{κ} ,则计算时间增加 m 倍 。即

$$T_{\mathrm{ime}}(m^{\kappa},1) = mT_{\mathrm{ime}}(\kappa,1)$$

定理 :在算法规模增加 m 倍的同时增加处理机数量 m 倍 ,若计算时间不变 ,则并行效率不变 。

证 :

$$T_{\mathrm{ime}}(m^{\kappa}, mP) = T_{\mathrm{ime}}(\kappa, P)$$

$$T_{\mathrm{ime}}(m^{\kappa},1) = mT_{\mathrm{ime}}(\kappa,1)$$

$$E_{\mathrm{fficent}}(m^{\kappa}, mP) = \frac{T_{\mathrm{ime}}(m^{\kappa},1)}{T_{\mathrm{ime}}(m^{\kappa}, mP)\ mp} = \frac{mT_{\mathrm{ime}}(\kappa,1)}{T_{\mathrm{ime}}(\kappa, P)\ mp}$$

$$= \frac{T_{\mathrm{ime}}(\kappa,1)}{T_{\mathrm{ime}}(\kappa, P)\ p} = E_{\mathrm{fficent}}(\kappa, P)$$

证毕 。

所以 ,可扩展性不变与计算时间不变等价 。

3　处理机计算速度和数量与通信 、I/ O 性能的关系

在式(1) 中假定计算时间 T_{ca}不变 ,并设处理机计算速度 $S_{\mathrm{ca}} = \frac{1}{t_{\mathrm{ca}}}$,有 :

$$T_{\mathrm{ca}} = \frac{TL^3}{\alpha d^4 n^2} t_{\mathrm{ca}} \tag{2}$$

可得 :

$$d = \left(\frac{TL^3}{\alpha T_{\mathrm{ca}}}\right)^{\frac{1}{4}} \left(\frac{1}{S_{\mathrm{ca}} P}\right)^{\frac{1}{4}} \tag{3}$$

假定延迟时间不变有 :

$$t_{\mathrm{la}} = \frac{\alpha T_{\mathrm{la}}}{4\ T} d \tag{4}$$

代入式(3)

$$t_{\mathrm{la}} = \frac{\alpha T_{\mathrm{la}}}{4} \left(\frac{L^3}{\alpha T_{\mathrm{ca}}\ T^3}\right)^{\frac{1}{4}} \left(\frac{1}{S_{\mathrm{ca}} P}\right)^{\frac{1}{4}} \tag{5}$$

假定通信时间不变 ,并设通信速度为 $S_{\mathrm{co}} = \frac{1}{t_{\mathrm{co}}}$有 :

$$T_{\mathrm{co}} = \frac{4e\ TL^2}{\alpha n d^3} t_{\mathrm{co}} \tag{6}$$

代入式(3) 可得 :

$$S_{\mathrm{co}} = \frac{4e}{T_{\mathrm{co}}} \left(\frac{TT_{\mathrm{ca}}^3}{\alpha L}\right)^{\frac{1}{4}} \left(S_{\mathrm{ca}}^3 P\right)^{\frac{1}{4}} \tag{7}$$

假定总 I/ O 时间不变 ,并设 $S_{\mathrm{co}} = \frac{1}{t_{\mathrm{io}}}$可得 :

$$(\frac{L}{d})^3 t_{io} = T_{io} \tag{8}$$

代入式(3)得

$$S_{io} = \frac{1}{T_{io}} (\frac{\alpha L T_{ca}}{T})^{\frac{3}{4}} (S_{ca} P)^{\frac{3}{4}} \tag{9}$$

由此可见,系统处理机的计算速度和数量与通信延迟时间和通信速率,I/O 速率之间存在相应的比例关系:通信速度 $S_{co} \propto O(S_{ca}^3 P)^{\frac{1}{4}}$,延迟时间 $t_{la} \propto O(\frac{1}{S_{ca} P})^{\frac{1}{4}}$,I/O 速度 $S_{co} \propto O(S_{ca} P)^{\frac{3}{4}}$。

4 计算与通信重叠的影响

如果每个处理机都由一个独立的通信管理硬件,可以实现通信与计算的操作重叠,这时,要求非边缘区域的计算时间大于等于通信时间,若延迟时间远远小于通信时间,边缘区域远远小于非边缘区域,即 $L \gg 2edn$,有:

$$\frac{L}{d} (\frac{L}{dn})^2 t_{ca} \geqslant \frac{4L^2 e}{d^2 n} t_{co}$$

带入式(3)得

$$S'_{co} \geqslant \frac{4e}{L} (\frac{TL^3}{\alpha T_{ca}})^{\frac{1}{4}} (S_{ca}^3 P)^{\frac{1}{4}} \tag{10}$$

其中 $S'_{co} = 1/t_{co}$为通信与计算重叠时的通信速度,由式(7)和(10)有

$$S'_{co} \geqslant \frac{T_{co}}{T_{ca}} S_{co} \tag{11}$$

对于多数应用课题,通信时间远远小于计算时间,即$\frac{T_{co}}{T_{ca}} \ll 1$,在通信与计算操作可以重叠的计算机上,只要通信时间不大于计算时间,即可将通信时间完全隐藏。所以对通信速度的最低要求是:

$$S'_{co} = \frac{T_{co}}{T_{ca}} S_{co} \ll S_{co} \tag{12}$$

与没有此项功能的计算机相比,对通信速度的要求可以适当降低。

5 试验结果与分析

我们以在数值预报模式中抽取的二阶扩散方程为例,在 IBM sp2 计算机上,对上述分析结果进行验证。取 $L = 130$ km, $T = 13$ s,分别用 $P_1 = 9$ 个与 $P_2 = 16$ 个处理机进行试验。并假定,增加处理机数量的目的是增加机算精度,在处理机数量为 $P_1 = 9$ 时,分辨率 $d_1 = 1$ km,根据式(2),有 $P_1 d_1^4 = P_2 d_2^4$ 可得 $d_2 = 0.866$,测量了计算、通信、I/O 时间。忽略延迟时间。由于在同一台计算机上,计算、通信、I/O 速度不变,根据前面的分析,我们通过测量 $P_1 = 9$ 时的通信时间,分别用式(6)和(8)推算增加分辨率后相应的通信、I/O

时间并与实测数据进行比较(见表 1)。可以看出二者是比较吻合的。证明分析是正确的。

表 1　不同分辨率的二阶扩散方程运算时间　　s

	分辨率(处理机数量)	
	1(9)	0.866(16)
计算时间	8.59	8.83 (8.59)
通信时间	0.2536	0.354 (0.39)
I/O 时间	0.9338	1.389 (1.43)

注:括号中的值是推算值。

6　结　论

本文通过建立初值问题的计算模型,分析了并行计算机系统的处理机速度,处理机数量与通信延迟时间、通信速度、I/O 速度之间的关系。指出对于解决该类问题的并行计算机有以下要求:

(1) 并行处理机的处理机数量、处理机速度、通信延迟、通信速度以及 I/O 速度形成有机的整体,提高处理机计算速度 S_{ca} 或增加处理机节点数量 P 的同时必须以 $O(S_{ca}^3 P)^{\frac{1}{4}}$ 的比例增加并行机的通信速度,以 $O(\frac{1}{S_{ca} P})^{\frac{1}{4}}$ 的比例降低延迟时间,以 $O(S_{ca} P)^{\frac{3}{4}}$ 的比例提高 I/O 速度。否则会降低系统的可扩展性。

(2) 计算与通信操作重叠,不仅可以提高并行效率,同时还有利于降低通信速度对系统可扩展性的影响。

参 考 文 献

1　刘德才,王鼎兴,沈美明,等. 数据并行的性能分析. 软件学报,1994,**5**(5):8~15.

2　Gustafson J L. Reevaluating Amdahl's Law. *Commun. ACM* (Association for Computing Machinery), 1988, **31**(5): 532~533.

3　Nussbaum D, Agarwal A. Scalability of parallel machine. *Commun. ACM*, 1991. **34**(3): 57~61.

4　Kai Hwang. 高等计算机系统结构. 王鼎兴,等译. 北京:清华大学出版社, 1995. 111~112.

SCALABILITY OF MASSIVELY DATA PARALLEL COMPUTING PROBLEMS

Jin Zhiyan

(*Chinese Academy of Meteorological Sciences, Beijing* 100081)

Wang Dinxing

(*Department of Computer Science, Tsinghua University, Beijing* 100084)

Abstract

The performance of distributed parallel systems is influenced by many factors, for exam-

ples, the total number of nodes in the system, node speed, I/O speed, latency and communication speed of the inter-network. Adding more nodes and/or using more powerful nodes can improve the performance, but I/O and communication could suffer from it. To determine the relationship between performance and those factors, the scalability of the massively parallel computers by using a data parallel model is analyzed. The relationship between the number of nodes, the speed of the nodes and the communication speed and latency of the links between nodes, the I/O speed of the system is obtained. The results shows that it is necessary to increase the I/O speed and the speed of the links (decrease the latency) by a certain ratio while increasing the number or the speed of the processors, so as to keep the scalability of the system. Only increasing the number or the speed of the processors will decrease the scalability of the system.

Key words: Parallel computing Scalability Data parallel processing