文章编号:1001 - 9081(2009)10 - 2849 - 03

基于前验负载差异的负载平衡性能模型

张理论,吴建平,宋君强
(国防科学技术大学 计算机学院,长沙 410073)
(zll0434@ sina. com)

摘 要:基于有限差分离散的并行应用非常普遍,针对此类问题的负载平衡性能评估,引入了一个刻画应用问题负载平衡能力的关键参数:最大负载变化率,推导了一个以并行效率为目标函数的负载平衡性能模型,涉及问题规模、并行通信计算比、离散格式复杂度和并行规模等。以 POP 全球海洋模式并行程序为测试实例,验证了该模型的性能。结果显示最大负载变化率作为衡量负载平衡程度的指标是有效的,基于模型的预测性能与实测性能在总体趋势上基本吻合。该性能模型对基于有限元、有限体积等其他局部离散格式的大型并行计算应用的负载平衡能力评估也具有参考价值。

关键词:数值并行计算;负载平衡;性能模型;负载变化率;通信计算比
中图分类号: TP393 **文献标志码**:A

Load-balancing evaluation model based on pre-defined variability

ZHANG Li-lun, WU Jian-ping, SONG Jun-qiang
(*College of Computer Science, National University of Defense Technology, Changsha Hunan* 410073, *China*)

Abstract: An evaluation model of load-balancing was presented for finite difference parallel computing. Maximal Load Variability (MLV) was introduced as a key index for load-balancing, and the quantitative model was constructed with parallel efficiency as the object function, involving ratio of communication to computing, degree of parallel, problem size and complexity of local numerical schemes. Both the parameter MLV and the model were verified by analyzing the POP global ocean circulation model benchmark. The results show that the performance derived from the evaluation model is consistent with that of wall-time measurement on the whole. And the model given in this paper is also useful for those parallel applications with local schemes such as finite element and finite volume.

Key words: numerical parallel computing; load balancing; evaluation model; load variability; ratio of communication to computing

0 引言

庞大的并行计算机系统,需要从应用软件角度建立结合体系结构和应用算法特点的性能模型。虽然目前已有多种并行计算模型和可扩展性模型[1-3],但专门针对负载平衡性能分析的较少。文献[4]作者提出了一个“负载平衡效率”概念,通过各并行任务墙钟时间总和与 P 倍最慢任务墙上时间之比来衡量并行程序的负载平衡能力;文献[5]作者对二维三温多物质非定常流体力学拉格朗日数值模拟程序,提出了一种多层均权负载平衡算法,并基于“负载平衡效率”做了理论分析。本文没有采用并行任务的墙钟时间作为后验度量,考虑到很多采用差分离散的并行应用其计算负载可以预估,基于并行任务负载的前验估计值建立了一个负载平衡性能模型。以数值模拟中的有限差分离散格式为研究对象,结合通信计算比、格式复杂度和并行计算规模等,从负载划分角度给出理想加速比和理想并行效率的表达式。通过引入最大负载变化率作为衡量负载平衡的性能指标,构建了一个基于前验负载差异的负载平衡性能评估模型。以典型的潜在千万亿次应用——POP 全球海洋模式为计算实例,对负载平衡性能模型进行了验证。

1 理想负载平衡的性能分析

1.1 问题假设和简化

时间发展方程的数值解在数值模拟中极为常见。在三维空间网格 $M \times N \times K$ 中,将问题加以简化限制:数值计算核心是基于预条件的 Krylov 迭代,矩阵矢量乘积是最基本操作;并行计算中只对水平方向进行两维数据剖分(如数值天气预报、海洋环流模拟等);并行计算任务是同构的,单个格点状态更新的计算量为 $w(i,j,k)$;并行计算子区域的伪边界宽度取 l,初始化和后处理串行时间不计。以上假设前提具有一般性:预条件 Krylov 迭代在时间发展方程的离散求解中占据主导地位;实际问题在垂直方向的物理特性(例如受重力作用或者地转作用等)导致数值模拟中垂直坐标设置复杂,垂直方向离散规模相对水平方向而言较小,因而水平两维剖分较为普遍。$w(i,j,k)$ 实质上是离散格式复杂度,对于两维并行剖分而言,这里只考虑水平方向的离散格式复杂度。一般情况下,$w(i,j,k)$ 在有限差分或有限元离散模板格式确定时取固定常数 $w = O(n)$,其中 n 为局部离散格式涉及的网格点数。对每个并行任务而言,计算负载和物理网格遵循简单的数值关系,不妨令 $M = N$,串行总负载为 $X = M^2Kw$,P 个并行任务的负载为:

收稿日期:2009 - 04 - 14;**修回日期**:2009 - 06 - 29。 **基金项目**:国家自然科学基金资助项目(40505023)。

作者简介:张理论(1975 -),男,河南南阳人,副研究员,博士,主要研究方向:高性能计算应用软件、大规模并行算法; 吴建平(1974 -),男,湖南新化人,副研究员,博士,主要研究方向:数值并行算法; 宋君强(1962 -),男,湖南宁乡人,研究员,博士生导师,主要研究方向:高性能计算、数值天气预报。

$x_i = \bar{x} + \Delta x_i;\ i = 1,\cdots,P$

其中:$\Delta x_i (i = 1,\cdots,P)$ 为负载变化量;负载平均值 $\bar{x}$(不妨假定 x,y 方向并行剖分分别为 $\sqrt{P}$,此时问题的通信计算比最小,区域角点不计)为:

$$\bar{x} = \left(M^2/P + 4lM/\sqrt{P}\right)Kw = (\eta^2 + 4\eta l)Kw;$$

$$\eta = M/\sqrt{P} \tag{1}$$

1.2 通信计算比、加速比和并行效率

考虑通信计算比 $C_P = t_{1P}/t_{2P}$(对于单个并行任务而言,一般由局部离散格式、并行度以及水平维度决定),其中 t_{1P},t_{2P} 分别为并行任务的通信和计算时间。对于确定的有限差分或有限元离散,理想情况下的 C_P 可预估如下:

$$C_P = \frac{t_{1P}}{t_{2P}} \approx \frac{\frac{4K\eta l}{10^9\Omega}}{\frac{\bar{x}}{(10^9\mu L)}} = \frac{4\mu LK\eta l}{\Omega\bar{x}} = \frac{4L\mu l}{\Omega(\eta + 4l)w} \tag{2}$$

其中 Ω、μ、L 分别为通信带宽(GB)、CPU 核主频(GHz)和 CPU 核的流水线部件数。例如某至强机群平台主要参数为:Ω = 0.7 GB,μ = 2.33 GHz,L = 4,对只需单个通信伪边界的五点差分格式,取 $l = 1, w = 5$。显然,通信计算比应满足 $C_P < 1$ 才能实现较好的并行性能。在固定平台上,通信计算比与并行计算规模、问题规模均密切相关。

引入理想并行负载下的理想加速比 λ_{ideal} 和理想并行效率 e_{ideal}:

$$\lambda_{ideal} = \frac{X}{\bar{x} + \bar{x}C_P} = \frac{M^2K}{(M^2K/P + 4KM/\sqrt{P})(1 + C_P)} = \frac{P}{(1 + 4/\eta)(1 + C_P)} \tag{3}$$

在负载足够大的情况下,理想加速比随着并行计算规模增加而近似线性增长:

$$e_{ideal} = \frac{1}{(1 + 4/\eta)(1 + C_P)} \tag{4}$$

2 负载平衡性能模型

实际计算中各个任务的负载通常不均衡,这往往是直接导致实际加速比小于理想加速比的重要因素。记 $\delta = \max \Delta x_i$ 最大负载变化量(假定同构任务和同构硬件节点,这里之所以没有采用均方根误差,主要考虑并行计算的时间往往是由最慢的并行任务决定)。为研究负载变化对并行计算的影响,引入最大负载变化率 α 衡量负载不平衡程度:$\alpha(P) = \delta/X$,α 显示相对于串行总负载规模的负载不平衡程度,直观上应满足:

$$0 < \alpha(P) < 1/P \tag{5}$$

将实际并行加速比和并行效率分别写为 $\alpha(P)$ 的函数,得出负载平衡的性能模型如下:

$$\begin{cases} \lambda_{real} = \dfrac{X}{\bar{x}(1 + C_P) + \delta} = \dfrac{1}{1/\lambda_{ideal} + \alpha} = \dfrac{\lambda_{ideal}}{1 + \alpha\lambda_{ideal}} \\ e_{real} = \dfrac{X}{\bar{x}(1 + C_P) + \delta}/P = \dfrac{\lambda_{ideal}}{1 + \alpha\lambda_{ideal}}/P = \dfrac{1}{(1 + 4/\eta)(1 + C_P) + \alpha P} \end{cases} \tag{6}$$

对式(6)在固定 M,P 和差分格式的条件下,图1给出了 α 变化时并行效率 e_{real} 的变化曲线。

3 负载平衡性能评估实例

POP 海洋环流模式是一个面向千万亿次计算的重要 Benchmark 程序,作为 NEC 地球模拟器、Cray 红色风暴、IBM 蓝基因等高端计算机的测试考题,其测试结果是各大厂商展示其机器性能的重要依据。POP 模式采用九点中心差分离散,基于预条件 GCR 迭代求解半隐式离散带来的椭圆方程,主要计算量集中在矩阵矢量乘积[6]。2003 年 12 月公布的 POP 2.0.1 版程序中包括两种数据剖分方式:传统笛卡尔分布和改进平衡分布。不同的数据剖分形式造成整体负载平衡程度差异很大,对程序性能影响显著。笛卡尔数据剖分采用简单两维拓扑分解,分别将 x,y 方向的任务数 N_x,N_y 和相应方向上的物理网格数作简单线性映射。一般各个物理格点上计算量比较一致,这种数据剖分在适当调整 N_x,N_y 后可以较好地实现负载平衡。在全球海洋模式中由于存在大量陆地点和岛屿,对应格点上的计算量可忽略不计,因而采用笛卡尔数据剖分往往会导致严重的负载不平衡。改进平衡数据剖分实质上是对笛卡尔剖分后的数据的进行子块再分,即在简单笛卡尔分布完成后,对并行剖分中的各个局部子块的计算量进行统计,并在 x,y 方向按照各个局部块的计算量大小作优先级重排序,依据优先级对以局部块为单位进行数据重分布,并尽可能保证局部块分布在基于笛卡尔分布的拓扑任务上或者只迁移到邻近任务上[6-7],这种子块再剖分思想类似于 GRAPES 格点大气模式的负载平衡设计[8]。

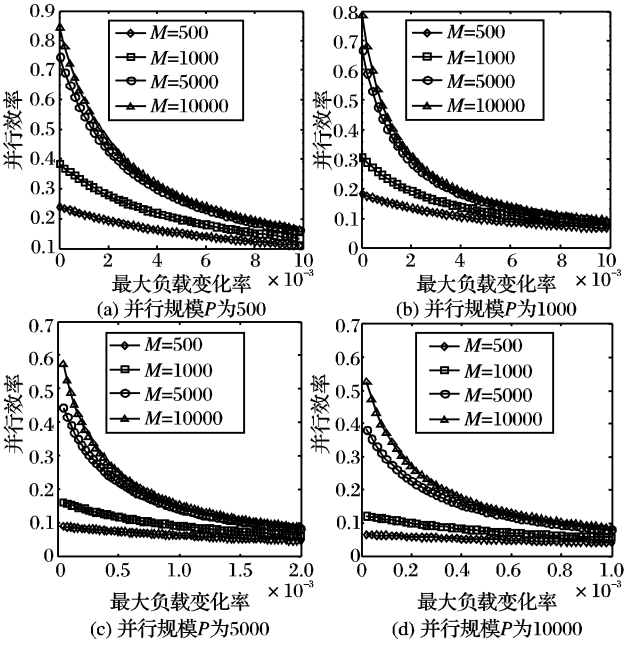


图1 并行效率和最大负载变化率 α 的关系

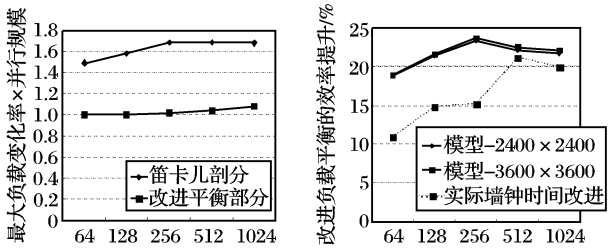
并行平台采用双路四核 Intel Xeon Clovertown/2.33 GHz,L1 cache 为 4×(32 KB 数据 + 32 KB 指令),L2 cache 为 8 MB,单 CPU 浮点性能 9.32 Gflops,互联采用 Voltaire Infiniband DDR,计算节点内存 4 GB,I/O 节点内存 8 GB,操作系统版本 Redhat AS4 2.6.18,Intel 编译器版本 10.0.023。POP 大规模算例:水平网格为 3 600×2 400,垂直方向 40 层,相当于全球 10 公里分辨率。海洋模式计算主要发生在海洋格点,落在陆地和岛屿的网格几乎不产生计算,且各个海洋格点计算量平均分布。因而可将各个并行任务所分配的海洋网格点数目作为其计算负载的度量指标,这种海洋网格点数目可在 POP 模式中插入代码进行统计,并由此获得最大负载变化率 α。表1给出了几种并行规模下不同剖分方式的负载平衡程度和性能改进。其中 A、B 分别表示简单笛卡尔剖分和改进平衡剖分;对每个 α,可由性能模型获得网格为 2 400×2 400 及 3 600×

3600两种离散情形的负载平衡并行效率,从而预测出当负载平衡改善 α 变化对 3600×2400 算例的效率改进近似值(容易知道 3600×2400 算例的效率介于 2400×2400 及 3600×3600两种网格离散情形之间)。

表 1　POP 模式负载平衡改进所带来的性能提升

并行规模	剖分方式	墙钟时间	α	αP	模型改进/%		墙钟时间改进/%
					2400×2400	3600×3600	
64	A	1067.82	1/43	1.488	18.79	18.91	10.96
	B	950.78	1/63	1.003			
128	A	463.47	1/81	1.580	21.45	21.64	14.85
	B	394.63	1/126	1.003			
256	A	220.94	1/152	1.684	23.36	23.64	15.23
	B	187.29	1/251	1.019			
512	A	98.09	1/304	1.684	22.13	22.50	21.19
	B	77.30	1/493	1.038			
1024	A	56.67	1/609	1.681	21.68	22.19	19.92
	B	45.34	1/947	1.081			

表 1 中的模型改进是指由最大负载变化率 α 的改进和性能模型预测出的效率提升。墙钟时间改进指改进平衡剖分相对笛卡尔剖分而言的性能提升程度,直接由墙钟时间推算获得,即 $\Delta e_r = (T_A - T_B)/T_A$,这里 T_A、T_B 分别为笛卡尔剖分和改进平衡剖分所对应的墙上时间。



(a) αP变化曲线　　(b) 并行效率的提升程度

图 2　POP 负载平衡改善对性能的影响效果

图 2 中:(a)为 αP 的变化曲线,显示"改进平衡"数据剖分明显降低了最大负载变化率;(b)为并行效率的提升程度,包括由性能模型预测的网格为分别为 2400×2400 和 3600×3600 算例的理论效率提升,以及由实际墙钟时间获得的并行效率提升。

这里 Δe_r 恰好是传统加速比意义下的并行效率改进。不妨假定最佳串行时间为 T_0,并行规模为 P,则有:

$$(T_A - T_B)/T_A = 1 - T_B/T_A = 1 - \frac{T_0/T_A}{T_0/T_B} = \frac{(T_0/T_B)/P - (T_0/T_A)/P}{(T_0/T_B)/P} = \frac{E_{BP} - E_{AP}}{E_{BP}} \tag{7}$$

这里 E_{AP}、E_{BP} 为传统意义下的并行效率。负载平衡性能改善后,从式(6)可以推出由模型预测的性能改进:

$$\frac{(\alpha_A - \alpha_B)P}{\left((1+4/\eta)(1+C_P) + \alpha_A P\right)\left((1+4/\eta)(1+C_P) + \alpha_B P\right)} \tag{8}$$

由表 1 和图 2(b)可以发现,对两种剖分方式,从 64 到 512 并行规模,程序均出现超线性加速,这很可能与 POP 模式设计中的缓存优化相关,从而使得由模型预测的效率提升幅度在 512 并行规模以内与由墙钟时间得到的性能改进存在较大差异。在 512 和 1024 并行规模,超线性加速现象消失,由模型预测的性能改进和实际墙钟时间的性能改进非常接近,且并行规模为 1024 时由模型预测和墙钟时间得到的性能改进均出现了较为显著的下滑。总的来说,由模型预测的性能改进与墙钟时间性能改进在总体趋势上是一致的,即先增后减,并在超线性加速现象消失的情况下非常接近。表 1 和图 2(a)显示尽管采用"改进平衡剖分"在很大程度上改善了程序性能,但仍不能满足 $0 < \alpha(P) < 1/P$,因而 POP 模式程序的负载平衡程度还有很大的优化空间。在并行可扩展性不受限时,改进负载平衡能力对整体性能的改善程度随着并行规模的增加更加显著,负载平衡程度对性能的影响越发敏感。

4　结语

对于涉及数千上万 CPU 核的超大规模并行计算,传统的加速比定义难以生效。本文提出的负载平衡性能模型,从负载分配角度对加速比和并行效率进行了诠释。目前并行计算领域存在众多性能模型,均将关键参数归结为墙钟时间作为后验度量,而本文的负载平衡性能模型采用了前验的负载差异估计量,这对很多基于格点计算的应用问题是可行的。应该指出,文中所给出负载平衡性能模型是做了理想简化的,例如假定计算同构,通过引入通信计算比假定计算负载和通信负载呈线性关系等。文中主要涉及计算负载,在性能模型中通过通信计算比将通信负载转换为计算负载。通过对 POP 模式的实例分析,验证了文中提出的负载性能参数——最大负载变化率和性能模型的可操作性及性能预测能力。

参考文献:

[1] 李晓梅,吴建平. 数值并行算法与软件[M]. 北京:科学出版社,2007.

[2] 迟利华,刘杰,胡庆丰. 数值并行计算可扩展性评价与测试[J]. 计算机研究与发展,2005,42(6):1073-1078.

[3] 徐小文,莫则尧. 并行代数多重网格算法可扩展性能分析[J]. 计算物理,2007,24(4):387-394.

[4] 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.

[5] 莫则尧. 一维高效动态负载平衡方法:多层均权法[J]. 计算机学报,2001,24(2):184-190.

[6] SMITH R D, GENT P. Reference manual for the parallel cean program (POP), Los Alamos Unclassified Report LA-UR-02-2484[R]. 2002.

[7] JONES P W, WORLEY P H, YOSHIDA Y, *et al*. Practical performance portability in the Parallel Ocean Program (POP)[J]. Concurrency and Computation: Practice and Experience, 2005,17(1):1317-1327.

[8] 伍湘君,金之雁,陈德辉,等. 中国新一代数值天气预报系统GRAPES模式的并行计算设计与实现[J]. 计算机研究与发展,2007,44(3):510-515.