

## 参数扫描应用的网格自适应调度

王庆江, 桂小林, 郑守淇

(西安交通大学电子与信息工程学院, 710049, 西安)

**摘要:** 提出一种自适应调度方法, 可使参数扫描应用在运行时保持近似的网格负载平衡. 为适应资源性能的动态性, 一个参数应用被分成若干作业顺序调度. 每个作业运行后反馈的网格负载信息用于调整下一个作业的调度, 使之适应资源性能波动, 从而使后一个作业在运行时实现网格负载的近似平衡. 每个作业被分成若干子作业, 分别指派到不同的网格资源. 子作业的运行时间构成一个网格负载向量, 从中可计算出网格负载失衡因子, 失衡因子表示作业运行时网格负载失衡的程度. 负载向量用于调整下一个作业的划分方法, 失衡因子用于调整下一个作业的规模. 较小的失衡因子可使作业的规模有更快的增长, 这样可使调度成本的增长速度慢于应用规模的增大速度. 实验表明, 自适应调度可保持近似的网格负载平衡, 与其他资源性能的静态调度相比, 可有效缩短参数应用的总运行时间.

**关键词:** 参数扫描应用; 自适应调度; 网格负载平衡; 网格负载向量; 网格负载失衡因子

**中图分类号:** TP393 **文献标识码:** A **文章编号:** 0253 - 987X(2004)02 - 0111 - 04

### Adaptive Scheduling for Parameter Sweep Applications in Grid

Wang Qingjiang, Gui Xiaolin, Zheng Shouqi

(School of Electronics and Information Engineering, Xi an Jiaotong University, Xi an 710049, China)

**Abstract:** A kind of adaptive scheduling was introduced to keep the approximate balance of grid load during parameter sweep applications' run. To adapt to the dynamic of resource capabilities, a parameter application was divided into some jobs scheduled sequentially. After each job finished its run, the feedback information about grid load was used to adjust the scheduling of next job to fit into the fluctuations of resource capacities, thus approximate grid load balancing could be implemented while the next job was running. Each job was divided into some sub-jobs which were assigned to different grid resources. According to sub-job execution time, a grid load vector was constructed, from which a grid load unbalance index indicating the unbalance extent of grid load during job's run was obtained. The load vector was used to adjust the partition of next job, and the unbalance index was used to change the scale of next job. The smaller indexes made job scale increase more rapidly, thus the increase of scheduling cost was slower than that of application scale. The experiments show that the adaptive scheduling can keep approximate balance of grid load, and can obviously shorten the total execution time of parameter applications in contrast with the static scheduling based on resource performance.

**Key words:** parameter sweep application; adaptive scheduling; grid load balancing; grid load vector; grid load unbalance index

网格上运行大规模科学仿真的计算资源以机群  
为主, 科学仿真可以划分为若干部分, 分别指派到不

同的机群. 参数扫描 (或称参数实验) 是许多科学领  
域都需要的, 即在一个多维参数空间上计算每一点

收稿日期: 2003 - 06 - 05. 作者简介: 王庆江 (1968 ~), 男, 博士生; 郑守淇 (联系人), 男, 教授, 博士生导师. 基金项目: 国家自然科学基金资助项目 (60273085); 国家十五“八六三”计划资助项目 (2001AA111081).

的一个目标函数. 参数扫描作为 Embarrassing 分布式应用,它包含很多彼此独立的同构的任务. 网格可以满足参数应用对计算资源的需求,但参数应用的运行性能还取决于怎样确定向各机群指派的任务个数,即怎样实现网格负载均衡.

任务彼此独立的分布式应用有许多调度算法<sup>[1]</sup>,这些算法在估算任务运行成本基础上指派任务,尽量实现资源负载均衡. 但是,网格是多用户共享、跨管理域的异构环境,资源性能动态性明显,对于运行时间长的作业,任务运行成本的预测往往失去意义,而且网格的资源模型不同于以往的异构环境,故上述算法难以借鉴. 本文提出一种自适应调度方法,将参数应用分成多个作业,在多次调度中完成,可更好地适应网格资源的动态变化,使各作业在运行时实现网格负载的近似平衡,从而缩短参数应用的运行时间. 文献[2,3]提出一种网格自适应调度算法,通过周期性测量网格性能,调整下一个作业的调度方法,旨在减少文件传输成本.

### 1 问题定义

假设网格包含  $k$  个机群,第  $j$  ( $[1, k]$ ) 机群的主机数为  $h_j$ ,则网格中的主机数可表示为向量,即  $H = (h_1, h_2, \dots, h_k)$ . 假设参数应用共包含  $x$  个任务,被划分为  $r$  个作业,第  $j$  ( $[1, r]$ ) 个作业的任务数为  $n_j$ ,则参数应用的任务数可表示为向量,即

$$N = (n_1, n_2, \dots, n_r), \quad n_j = x$$

网格调度分为两层<sup>[4]</sup>,上层为一个网格调度器,负责把一个作业划分成若干子作业,不同子作业被指派到不同机群;下层为若干机群调度器,负责为子作业分配资源. 网格调度适应网格资源性能差别的关键是如何根据机群性能划分作业. 第  $i$  个作业的划分可表示为向量,即

$$J_i = (m_1, m_2, \dots, m_k), \quad m_j = n_i$$

在第  $i$  次调度下,各机群完成各自的子作业. 假设第  $j$  ( $[1, k]$ ) 机群完成子作业的时间为  $t_j$ ,则网格完成第  $i$  个作业的时间可用向量  $T_i = (t_1, t_2, \dots, t_k)$  表示.

定义1 假设网格完成第  $i$  个作业的时间向量为

$$T_i = (t_1, t_2, \dots, t_k), \quad t = \frac{1}{k} \sum_{j=1}^k t_j$$

则  $(\frac{t_1}{t}, \frac{t_2}{t}, \dots, \frac{t_k}{t})$  称为第  $i$  次调度下的网格负载向量,记为  $L_i$ .

当  $t_j$  ( $[1, k]$ ) 彼此相等时,  $L_i = (1, 1, \dots, 1)$  表示第  $i$  次调度实现了网格负载平衡. 假设  $L_i$  的第  $j$  分量为  $l_{i,j}$ ,  $l_{i,j}$  反映了第  $j$  机群参与运行第  $i$  个作业时的负载大小,  $l_{i,j} > 1$  表示机群负载偏重,  $l_{i,j} < 1$  表示机群负载偏轻,  $l_{i,j} = 1$  表示机群负载适中.

定义2 假设第  $i$  次调度下网格负载向量为

$$L_i = (l_1, l_2, \dots, l_k), \quad j = \begin{cases} l_j, & l_j = 1 \\ 1/l_j, & l_j < 1 \end{cases}$$

则  $E(\cdot)$  表示  $j$  ( $[1, k]$ ) 的均值,  $E(\cdot) - 1$  称为第  $i$  次调度下的网格负载失衡因子,记为  $\delta_i$ .

当  $L_i = (1, 1, \dots, 1)$  时,  $\delta_i = 0$ .  $L_i$  中各分量之间差别越大,  $\delta_i$  越大,故  $\delta_i$  是对网格负载失衡程度的量化.

定义3 假设参数应用被分为  $r$  个作业,第  $i$  个作业的完成时间为  $t_i$ , 网格负载失衡因子为  $\delta_i$ ,  $t = \frac{1}{r} \sum_{i=1}^r t_i$ , 则  $\sum_{i=1}^r \delta_i t_i / t$  称为按时间平均的网格负载失衡因子,记为  $\bar{\delta}$ .

反映参数应用运行中网格负载失衡的程度. 自适应调度就是使  $|\bar{\delta}| > 1$  自动适应网格资源性能的动态变化,使  $\bar{\delta}$  尽可能小,以缩短参数应用的总运行时间.

### 2 自适应调度方法

定义4 假设第  $j$  个作业被分成  $k$  个子作业,分别由  $k$  个机群运行,第  $i$  ( $[1, k]$ ) 机群中每个主机平均得分任务的数目为  $v_i$ ,  $\bar{v} = \frac{1}{k} \sum_{i=1}^k v_i$ , 则向量  $(\frac{v_1}{\bar{v}}, \frac{v_2}{\bar{v}}, \dots, \frac{v_k}{\bar{v}})$  称为第  $j$  个作业的网格任务数向量,记为  $Q_j$ .

图1是一个自适应调度模型,自适应调度可使作业划分适应动态变化的网格资源性能,实现网格负载均衡,并尽可能减少调度次数.

图1中,  $Q_0 = (1, 1, \dots, 1)$  表示向网格中各主机指派相等数目的任务,  $T_i$  是网格完成第  $i$  个作业的

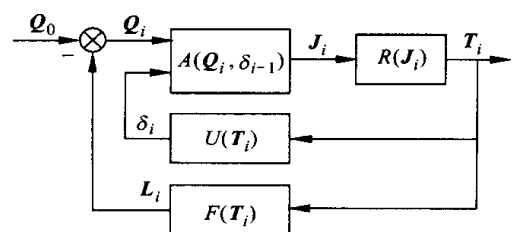


图1 自适应调度模型

时间向量,  $F$  将  $T_i$  映射为网格负载向量  $L_i$ ,  $U$  根据  $T_i$  计算第  $i$  次调度下的网格负载失衡因子  $\alpha_i$ ,  $Q_i$  为第  $i$  个作业的网格任务数向量. 假设  $Q_i, L_{i-1}, Q_{i-1}$  和  $Q_i$  的第  $j$  分量分别为  $q_{0,j}, l_{i-1,j}, q_{i-1,j}$  和  $q_{i,j}$ , 则  $q_{i,j} = \frac{q_{0,j}q_{i-1,j}}{l_{i-1,j}}$ , 即  $q_{i,j} = \frac{q_{i-1,j}}{l_{i-1,j}}$ . 负反馈使  $Q_i$  趋于稳定,  $L_i$  趋于  $(1, 1, \dots, 1)$ . 规定:  $L_0 = (1, 1, \dots, 1)$ .

第  $i$  次调度的作业划分向量为

$$J_i = A(Q_i, \alpha_{i-1}) = \begin{cases} Q_i \cdot J_{i-1}, & \alpha_{i-1} = 1 \\ (Q_i \cdot J_{i-1}) / (\alpha_{i-1}), & \alpha_{i-1} < 1, 0 < \alpha_{i-1} < 1/\alpha_{i-1} \end{cases} \quad (1)$$

其中,  $\alpha_{i-1}$  决定第  $i$  个作业的规模是否在第  $i-1$  个作业的规模基础上有所增大, 以及增大了多少. 这里以  $\alpha_{i-1} = 1$  为是否增大作业规模的界限, 当  $\alpha_{i-1} = 1$  时, 认为上次调度的网格负载失衡程度较重, 不增大作业规模; 当  $\alpha_{i-1} < 1$  时, 认为网格负载失衡程度较轻, 可适当加大作业的任务数. 增大作业规模的方法有多种, 如果规模增长过快, 每个作业的运行时间变长, 适应资源性能变化的能力减弱. 反之, 作业运行时间变短, 作业数变多, 适应网格资源变化的能力增强. 作业规模增大的方法取决于网格资源性能的稳定性, 式(1)中采用除以  $\alpha_{i-1}$  的方法. 网格资源性能越稳定,  $\alpha$  的取值越小. 规定:  $\alpha_0 = 1$ .

为获得  $T_1$ , 在机群性能差别未知的情况下, 向每个机群指派的任务应尽可能少, 但提供机群任务数一定的下调空间也是必要的, 故规定  $J_0 = (2h_1, 2h_2, \dots, 2h_k)$ ,  $h_j$  是第  $j$  ( $j = 1, k$ ) 机群的主机数.

$R$  不是一一映射函数, 表示实际的网格资源, 在  $J_i$  下资源完成作业的时间向量为  $T_i$ . 一般这是不可重复的过程, 因为网格资源性能是动态变化的.

### 3 实验数据分析

下面的实验将验证自适应调度能否有效实现参数应用运行中的网格负载平衡, 从而缩短参数应用的运行时间. 网格实验床<sup>[5]</sup>包括 3 个机群, 见表 1.

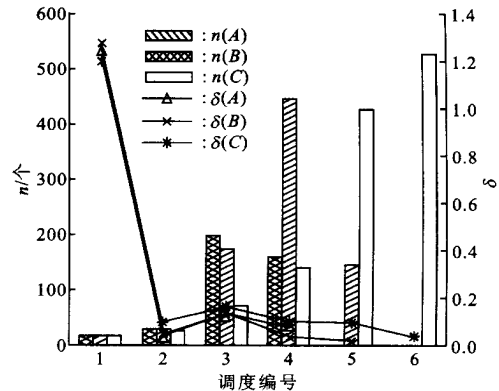
表 1 网格实验床中的机群

机群编号	主机 CPU	主机主频/MHz	主机数
1	PowerPC-POWER3	375	4
2	Intel P4	2 000	2
3	UltraSPARC	167	3

由于未找到实际的参数应用, 这里以 NPB<sup>[6]</sup> 中的 EP(每任务计算 335 544 个随机数) 作为参数应

用的任务. 假设有 3 个参数应用, 分别包含 400、800 和 1 200 个这样的任务.

对于自适应调度, 网格主机数向量  $H = (4, 2, 3)$ ,  $J_0 = (8, 4, 6)$ , 分别在网格实验床上运行 3 个参数应用. 网格实验床的用户很少, 网格资源性能稳定, 令  $\alpha = 3$ . 在自适应调度下, 3 个参数应用在调度次数、作业规模和网格负载失衡因子上的比较如图 2 所示.



A、B 和 C 分别表示 400、800 和 1 200 个任务的参数应用;  $\alpha$  为网格负载失衡因子;  $n$  是作业规模

图 2 3 个参数应用的自适应调度比较

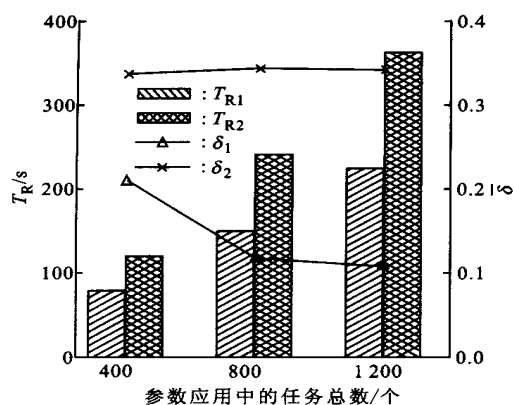
从图 2 看, 第 1 个作业在运行时网格负载明显不平衡, 但以后的作业都基本实现了网格负载平衡. 在网格负载平衡时, 作业包含的任务数逐步变大(最后一个作业由于包含剩余的任务, 故任务数未必大). 在参数应用的规模增大时, 调度次数增长缓慢. 自适应调度可以在没有任务运行成本预测时, 使绝大多数任务运行于网格负载平衡状态.

基于资源性能的静态调度需要估算资源性能, 这里用一任务  $S$  类 EP 的 benchmark 值的倒数作为各机群的主机性能, 分别为 0.054 4、0.081 3 和 0.028 7. 基于资源性能的静态调度只需一次作业划分, 根据各机群主机数、主机性能, 在网格负载平衡的目标下, 包含 400、800 和 1 200 个任务时的作业划分向量分别是  $J = (185, 140, 75)$ 、 $J = (370, 280, 150)$  和  $J = (555, 420, 225)$ . 在网格实验床上运行 3 个参数应用, 记录各机群完成子作业的时间和参数应用的总运行时间, 计算网格负载失衡因子.

图 3 是两种调度在按时间平均的网格负载平衡因子和参数应用总运行时间上的比较.

从图 3 可见, 与基于资源性能的静态调度相比, 自适应调度实现了更好的网格负载平衡, 使参数应用的运行时间更短. 参数应用的规模越大, 自适应调

度的优越性越明显。



1,2 分别表示自适应调度和基于资源性能的静态调度;  $T_R$  表示参数应用的总运行时间;  $\delta$  为按时间平均的网格负载失衡因子

图3 自适应调度和基于资源性能的静态调度比较

## 4 结论

由于网格资源的动态变化,基于资源性能的静态调度不能取得很好的网格负载平衡.本文为参数应用提出了网格中的一种自适应调度方法,它将参数应用划分为多个作业,在多次调度中完成参数应用的运行,从而能更好地适应网格资源性能的频繁波动.每个作业被划分为若干子作业,分别由不同的网格资源完成.网格负载向量反映一个作业运行时各网格资源的负载轻重,用于调整下次调度时的作业划分方法.网格负载失衡因子反映一次调度下网格负载的失衡程度,用于确定下一个作业的规模是否扩大,以及扩大了多少.失衡因子有助于减少调度

成本.实验表明,自适应调度使参数应用在运行时实现了近似的网格负载平衡,比基于资源性能的静态调度有更短的应用运行时间.

## 参考文献:

- [1] Maheswaran M, Ali S, Siegel HJ, et al. Dynamic matching and scheduling of a class of independent tasks onto heterogeneous computing systems [A]. 8th Heterogeneous Computing Workshop, San Juan, Puerto Rico, 1999.
- [2] Casanova H, Legrand A, Zagorodnov D, et al. Heuristics for scheduling parameter sweep applications in grid environments [A]. 9th Heterogeneous Computing Workshop, Cancun, Mexico, 2000.
- [3] Casanova H, Obertelli G, Berman F, et al. The AppLeS parameter sweep template: user-level middleware for the grid [A]. The 2000 ACM/ IEEE conference on Supercomputing (CDROM), Dallas, USA, 2000.
- [4] Schwiegelshohn U, Yahyapour R. Attributes for communication between scheduling instances [EB/OL]. <http://www.gridforum.org/Documents/GFD/GFD-F-6.pdf>, 2002-12-17.
- [5] 桂小林, 钱德沛, 何戈. 基于校园网络的元计算实验系统 WADE 的设计与实现 [J]. 计算机研究与发展, 2002, 39(7): 888~894.
- [6] NASA Advanced Supercomputing. NAS parallel benchmarks [EB/OL]. <http://www.nas.nasa.gov/Software/NPB>, 2002-11-19.

(编辑 苗凌)

## 陶文铨教授担任两个国际期刊的中国区副主编

受国际传热传质权威期刊《International Journal of Heat and Mass Transfer》和《International Communications in Heat and Mass Transfer》的执行主编、美国伊利诺斯大学机械系 Minkowycz 教授和 Hartnett 教授,以及 Pergamon 出版社的委托,从 2004 年 1 月起,我校陶文铨教授担任这 2 种期刊的副主编,负责中国地区稿件的评审与接收工作.这是陶文铨教授继去年担任国际著名传热传质期刊之一《Numerical Heat Transfer》的顾问编委之后,在国际传热传质界获得的又一殊荣.这也表明西安交通大学的传热学研究进一步得到国际学术界的认可.