

文章编号: 1005—0523(2009)02—0081—05

# 基于动态竞标机制的密集计算网格作业调度模型

柴亚辉<sup>1,2</sup>, 周 娟<sup>1</sup>, 刘觉夫<sup>1</sup>, 胡林峰<sup>1</sup>

(1 华东交通大学 信息工程学院, 江西 南昌 330013; 2 上海大学 计算机工程与科学学院, 上海 200072)

**摘要:**通过对当前密集计算网格上的作业调度系统的不足之处进行了分析, 继而提出将经济模型引入到网格的作业调度系统, 并提出了一个基于动态竞标机制的分级网格作业调度策略。最后, 通过在高性能集群网格节点自强 3000 与自强 2000 上的模拟网格资源上进行了对比实验。

**关 键 词:**密集计算网格; 动态竞标机制; 作业调度; 价格策略

**中图分类号:** TP393.01

**文献标识码:** A

密集计算网格主要是在进行问题求解时产生很大的存储和计算需求, 对参与求解的计算机的计算性能有很高的要求, 因此构建密集计算网格系统的节点往往以集群、服务器等高性能计算机为主。而网络技术能够十分有效地提高计算的吞吐率, 将大量空闲的计算机的计算资源集中起来, 提供给对计算性能要求很高而在单机或者几台计算机上很难短时间解决的问题<sup>[1]</sup>。网格作业调度的主要目的与作用是根据不同的网格作业任务需求与当前的网格资源状况进行资源分配, 实现对网格资源的优化使用, 为用户提高更好的服务质量。

当前密集计算网格的调度系统中主要采用的是性能驱动型的网格作业调度策略, 主要是基于优先级的启发式调度, 一般采用基于优先级的“双最”调度策略(即最高优先级作业运行在满足其资源需求的最高性能资源节点上), 但该策略有以下不足: (1) 单靠静态的优先级显然不能反映出各资源性能的动态变化情况; (2) 只是按照作业优先级和资源优先级进行调度, 有时会导致调度的资源并不一定是消费者最需要的或最适合的; (3) 在某些特定情况下, 会造成某些低优先级作业的饿死现象。

## 1 动态竞标经济模型

基于当前密集计算网格系统在实际运行中的不足, 我们将经济型调度模型引入网格作业调度系统。经济学尤其是市场机制非常适合解决计算网格环境下的作业调度和资源分配问题。网格是一个异构的、动态的分布式环境, 对资源的使用和资源的供应都是在不断的变化之中。通过引入经济学, 特别是基于价格的市场机制, 由价格的浮动来反映资源供需情况的动态变化<sup>[1]</sup>, 通过供需均衡实现资源优化配置, 能够很好地反映网格的动态特征。本文提出一个基于 OGSA 框架的新的经济学调度模型, 并利用 Globus Toolkit 网格中间件和 PBS Pro 作业管理排队系统, 将各种网格资源封装成“网格服务”, 并且具有相应的动态

收稿日期: 2009—01—20

基金项目: 南昌市重点科技攻关项目(洪财企[2008]68号)

作者简介: 柴亚辉(1977—), 男, 河北高阳人, 讲师, 研究方向为网格计算、高性能计算技术。

价格<sup>[2]</sup>。为了使供需双方都获得最大的受益,在网格的全局调度中采用动态价格竞标策略对网格作业所需的资源进行调度分配。

2 网络作业调度算法设计

2.1 两级调度结构

在网格计算环境中,作业调度实际上分为两级,全局网络作业调度和本地资源上的本地作业调度<sup>[4]</sup>。当用户递交作业后,由系统选择一个合适的资源进行实际运算。图 1 为一个基于两级调度的密集计算网络作业系统调度状态流程图。在全局调度层,采用动态竞标机制的经济模型调度策略,而在本地调度中采用性能型的作业调度策略。全局调度包括两个部份,一个是预测即将调度的网络作业在各个资源服务上运行的时间,寻找满足作业时间约束的资源,称为作业运行时间预测;另一个是向满足作业时间约束的资源服务满足预算条件的资源中进行竞价,选择运行费用最低的资源,即动态竞标。

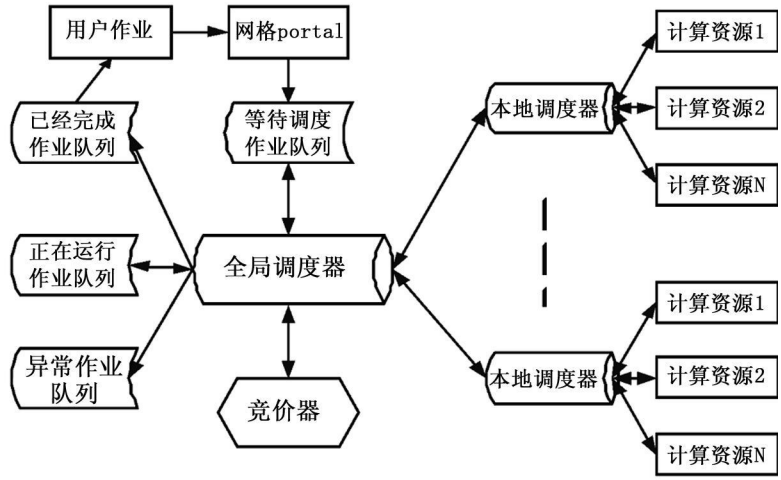


图 1 网络作业调度状态流程图

2.2 网络作业约束条件

作业执行时间的预测(或者说作业计算量的预测)是经济型网络调度策略的基础,它决定了经济型网络全局调度的成功率<sup>[3]</sup>。为了减轻用户的负担,本文采用基于历史数据的预测模型,同时为了兼顾算法准确性与快速性的平衡,采用平均值与标准差算法: $\delta$ 表示作业历史数据,预测值  $\text{prediction} = \min(\max(\delta), \text{mean}(\delta)) + 3 \cdot \text{stdev}(\delta)$ 。以上预测的作业执行时间并没有考虑到资源相互间的差异性,而实际上,各种资源的性能不同,平台各异,因此相同的作业在不同资源平台上执行的时间差别很大,还需要对抽象执行时间结合各个资源的特点进行调整,以求更准确地估算出作业实际的执行时间。在已知各资源性能系数的情况下,性能较好的资源,执行时间就短一些,性能较差的资源,执行时间就长一些。但此时间还只是估算的时间,我们以此时间进行调度还有较大的风险,可能仍会超出截止期限,因此还要在此时间上再加上一个误差系数  $\beta$  以最大限度的确保调度质量<sup>[5]</sup>。公式如下:  $T = T^* \epsilon + T^* \beta$  其中  $\epsilon$  是资源的性能系数。

2.3 动态竞标作业调度算法

用户通过网络 Portal 的作业提交页面可以向网格提交多个作业,每提交一个作业都由作业生成器生成该作业的作业节点(JobNode),保存作业的需求信息,并将其发送到各个资源服务端,由资源服务端的本地调度器针对本次作业进行本地资源的资源匹配操作,若满足要求,本地调度器则向全局调度器出价(竞

标),然后全局调度器再将本次作业委托与满足作业预算约束条件的出价最低者(竞标成功者)调度执行。如果没有满足条件的资源,则由用户选择进入等待队列等待下一轮调度,还是修改作业的需求和约束条件后重新提交。

全局网格作业调度决策函数可以表示如下:

$$F_1(J, T, R) = \begin{cases} P(J, T, R, Q_w) & \text{if cond}^1 \\ F_1(J, T_{new}, R_{new}) & \text{if cond}^2 \\ S(j, T, R, Q_r) & \text{if cond}^3 \\ T(J, T, R, Q_r) & \text{if cond}^4 \end{cases}$$

本地资源作业调度决策函数可以表示为如下:

$$F_2(J, T, R) = \begin{cases} M(J, R, T_b, R_l) & \text{if cond}^5 \\ R(J, T, R, Q_r) & \text{if cond}^6 \\ P(J, T, R, Q_{w1}) & \text{if cond}^7 \end{cases}$$

其中 Cond<sup>1</sup>表示:  $\{J \cdot req > R_l \cdot all \cup J \cdot req > R_l \cdot free\} \cap R \cdot flag = false$

Cond<sup>2</sup>表示:  $\{J \cdot req > R_l \cdot all \cup J \cdot req > R_l \cdot free\} \cap R \cdot flag = true$

Cond<sup>3</sup>表示:  $J \cdot req < R_l \cdot free \cap J \cdot bid \leq R_l \cdot price$

Cond<sup>4</sup>表示:  $Q_r \neq null \cap J \cdot req < R_l \cdot free$

Cond<sup>5</sup>表示:  $R < R_l$

Cond<sup>6</sup>表示:  $Q_r \neq FULL \cap J \cdot bid \leq R_l \cdot price \cap R \leq R_l \cdot free$

Cond<sup>7</sup>表示:  $Q_r = FULL \cup R > R_l \cdot free$

函数  $P(J, T, R, Q_w)$ :调度子函数,表示将当前的作业节点  $J$  加入到等待调度的作业队列  $Q_w$  中去。

函数  $F_1(J, T_{new}, R_{new})$ :调度决策函数,表示对修改作业需求  $R_{new}$  与约束条件  $T_{new}$  后的用户作业重新进行调度决策。

函数  $S(J, T, R, Q_r)$ :调度决策函数,表示将本次作业委托与满足作业预算约束条件的出价最低者(竞标成功者)调度执行,并将其加入到正在运行作业队列  $Q_r$  中。

函数  $T(J, T, R, Q_r)$ :调度决策函数,表示将本次作业将其发送到各个资源服务端,由资源服务端的本地调度器针对本次作业进行本地资源的资源匹配操作,并将其加入本地资源匹配队列  $Q_r$  中。

函数  $M(T, R, T_b, R_l)$ :调度决策子函数,表示将从全局调度器中传过来的作业需求  $R$  与约束条件  $T$  在本地资源进行资源匹配,并综合本地资源条件  $R_l$  与本地约束条件给出一个竞价,并将该竞价值回送给全局调度器。

函数  $R(J, T, R, Q_{rl})$ :调度决策子函数,表示将从全局调度器发送过来的作业调度到本到的资源节点上运行,并将其加入到本地运行作业队列中  $Q_{rl}$  中。

函数  $P(J, T, R, Q_{wl})$ :调度子函数,表示在本地资源调度器将当前的作业节点  $J$  加入到本地等到调度的作业队列  $Q_{wl}$  中。

### 3 网格本地资源调度出价策略

网格环境下资源是动态的,本地资源上可用结点个数、计算能力,以及作业的资源需求等都是随时变动的<sup>[5]</sup>。故面对  $N$  个计算结点,对  $M$  个作业(每个作业需求  $P$  个结点,  $T$  个运行时间)进行服务质量量化时,使总的作业吞吐率最大,并且需要满足各个作业不同的  $deadline$ (其中,每个参数都是随机变动的),实际上是个 NP 非确定性问题。因此在设计系统的时候,采用的是基于资源供需关系和应用时限相结合的动态价格策略,具体做法为按照当前作业的时限比和当前资源的负载情况来动态生成价格。所涉及的几

个价格变量为：

- s-Price 资源当前的标准价格
- m-price 资源的最低价格(固定值)
- b-price 资源的基准价格(固定值)
- price 资源当前的价格。

当前的价格 price是随着作业的执行时间 T和 deadline之比而变化的,公式计算为

$$price=\mu*(T/deadline)*s-Price$$

同时当前的标准价格也是随着当前资源的负载而变化的,计算公式为

$$s-Price=\nu*(load_{new}/load_{ideal})*b-Price$$

其中： $\mu$ 为公式的调整系数， $load_{new}$ 为资源当前的负载， $load_{ideal}$ 为资源期望的理想负载。同时，资源的当前价格不能小于资源的最低价格 m-price 否则取最低价格。

4 系统性能比较测试

统测试采用 4个小型的分布式集群计算资源,分别取自上海大学集群式高性能计算机自强 3000和自强 2000的部分结点。根据测试机器硬件的性能参数,并结合我们实际需要,特指定资源的性能系数、基准价格及理想负载见表 1。

表 1 模拟的网格测试资源

| 计算资源 | 性能系数 | 基准价 | 最低价 | 理想负载 /% |
|------|------|-----|-----|---------|
| R1   | 1    | 10  | 10  | 89      |
| R2   | 0.9  | 5   | 40  | 78      |
| R3   | 0.5  | 15  | 5   | 75      |
| R4   | 0.3  | 8   | 5   | 60      |

在实验中,将本文调度策略与基于优先级的启发式调度<sup>[6]</sup>,即基于优先级的“双最”调度策略进行了比较。我们以作业的数量为横坐标,系统的利用率为纵坐标,在基于理想性能比的基础上对两种算法进行比较。

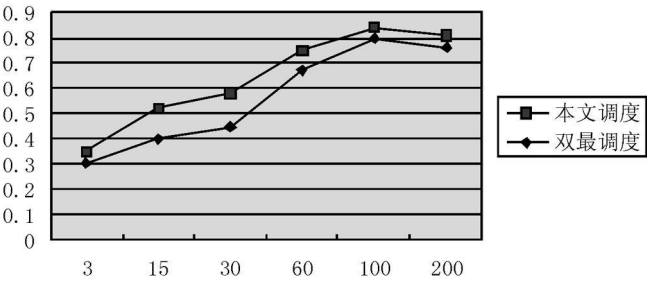


图 2 两种调度算法的用户资源利用率比较

从图 2中可以看出,本文所提出的基于动态价格竞标策略的调度算法可以在基于理想性能比的基础上进一步提高了网格用户资源的利用率。

5 结束语

在网格环境中,大量地理上分布的资源为不同的组织所拥有,这些组织具有不同的使用规则、不同的计费模型、不一样的负荷能力和不同的使用模型。因此,传统的资源管理和调度方式在网格中并不完全适

用,在网格环境中使用经济学原理进行资源的管理和调度是网格计算从理论走向实际应用所跨越的突破性的一步。针对分布式的资源特点分为两端:中心式的全局调度端和随资源分布的资源服务端。全局调度端负责用户的管理和全局作业的调度;资源服务端负责对资源的管理、本地作业的调度和作业的监控。作业约束条件和资源的报价能有效影响用户和资源提供者双方的利益,最终达到均衡,并能充分体现出资源性价比的优势,优化网格资源市场的配置,充分提高了网格资源的利用率。

## 参考文献:

- [1] Huang Changqin, Chen Deren, Zheng Yao, et al. Performance-driven task and data co-scheduling algorithms for data-intensive applications in grid computing [A]. YU J X, LIN X, LU H, et al. Proceedings of the APWeb 2004, lecture notes in Computer Science[C]. Berlin: Springer-Verlag, 2004.
- [2] Buyya R, Abramson D, Giddy J. An economy driven resource management architecture for global computational power grids [A]. 2000 International conference on parallel and distributed processing techniques and applications[C]. Las Vegas: PDPTA, 2000.
- [3] Buyya R. Economic-based distributed resource management and scheduling for grid computing[EB/OL]. <http://www.buyya.com/thesis/>. PhD Thesis, Monash University, Australia, 2002-04-12.
- [4] 柴亚辉,李洪刚,顾训穰. 基于资源角色分类的密集计算网格作业管理[J]. 计算机应用与软件, 2006, 23(7): 37-38.
- [5] 沈新超,郑衍衡. 一个基于全局竞标机制的网格调度系统[J]. 计算机应用与软件, 2008, 23(7): 209-211.
- [6] 张伟哲,方滨兴,胡铭曾,张宏莉. 基于信任 QoS增强的网格服务调度算法[J]. 计算机学报, 2006, 29(7): 1157-1165.

# Job Scheduling Model Designed for Intensive Computational Grid on the Basis of Dynamic Bidding Mechanism

CHAI Ya-hui<sup>1,2</sup>, ZHOU Juan<sup>1</sup>, LIU Jue-fu<sup>1</sup>, HU Lin-feng<sup>1</sup>

(1. School of Information Engineering, East China Jiaotong University, Nanchang 330013, China; 2. School of Computer Engineering and Science, Shanghai University, Shanghai 20072, China)

**Abstract:** Based on the analysis of deficiency of the current grid job scheduling system in intensive computational grid, this article introduces the economic model into grid job scheduling system, and proposes a graded grid job schedule strategy on the basis of a dynamic bidding mechanism. A grid scheduling system based on global bid mechanisms is proposed and implemented. The comparative experiment is conducted through the grid at high-performance cluster nodes of simulated grid resources of Ziqiang 3000 and Ziqiang 2000.

**Key words:** intensive computational grid; dynamic bidding mechanism; job scheduling price strategy

(责任编辑:王建华)