第 33 卷 第 3 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.3 Computer Engineering February 2007
·软件技术与数据库· 文章编号：1000—3428(2007)03—0076—03 文献标识码：A 中图分类号：TP311

计算网格工作负荷的建模

王庆江[1]，张 琳[2]
（1. 中国海洋大学计算机科学系，青岛 266071；2. 河南财经学院计算中心，郑州 450002）

摘 要：为评估计算网格中的作业调度，建立了网格工作负荷模型。在不同的节点，作业的运行时间不同；在不同的节点之间，作业的迁移开销不同。定义了不依赖网格资源性能的纯运行时间和纯迁移开销。借鉴并行计算机的工作负荷模型，可得到并行度、纯运行时间和到达间隔的分布。构建了作业提交位置、纯迁移开销、纯运行时间估计因子、完成期限的分布。应用实例表明，由网格工作负荷模型可获得 各种工作负荷，支持对作业调度的全面评估。

关键词：计算网格；作业调度；工作负荷模型

Workload Modeling of Computational Grids

WANG Qingjiang[1], ZHANG Lin[2]
(1. Department of Computer Science, Ocean University of China, Qingdao 266071;
2. Computer Center, Henan Institute of Finance & Economics, Zhengzhou 450002)

【Abstract】 To evaluate job schedule on computational grids, the model of grid workload is constructed. On different nodes, job runtimes are different. Between different nodes, the costs of job migrations are different. Thus, the pure runtime and the pure migration cost are defined, which are independent of the performances of grid resources. The distributions of parallel degree, pure runtime, and arrival interval are obtained from workload models of parallel computers. Besides, the distributions of submittal location, pure migration cost, estimate factor of pure runtime, and deadline are constructed. Application instances show various workloads can be obtained from the model of grid workload to support comprehensive evaluation of job scheduling.

【Key words】 Computational grids; Job schedule; Workload model

基于离散事件的调度模拟是评估网格作业调度的有效方法。事件一般有作业提交、作业开始运行、作业运行结束等。事件发生时，触发相应的网格操作。由一个虚拟时钟模拟墙上的时钟，每过一个时间单位，检查是否有事件发生。某个（些）作业完成时，模拟实验结束。Simgrid 和 GridSim 都是这类系统。

网格工作负荷是指一段时期里网格收到并完成的一系列作业，每个作业的参数包括提交时刻、提交到哪个网格调度器、并行度、在某个计算系统（又称节点）上运行时的运行时间，有时还包括作业的完成期限、在某种网络性能下的作业迁移开销等，网格工作负荷可用作调度模拟的输入。

目前难以获得真实的网格工作负荷，而且，真实的工作负荷往往不具代表性，很难用于各种网格配置下的实验。对网格工作负荷建模，通过调整模型参数，可以产生所需的工作负荷。本文通过扩展并行计算机的工作负荷模型，为网格建立了工作负荷模型。

1 相关研究

对于一个并行计算机，如IBM SP2，工作负荷记录在作业跟踪（trace）中。作业跟踪一般是一个ASCII文件，每一行代表一个作业，包含作业的并行度、运行时间、到达时刻等。用作业跟踪驱动模拟实验，可“真实”测试系统的工作。作业跟踪是特定系统在某段时期的工作情况，故模拟实验的结果未必适用其他系统或负荷情况。用模型产生的工作负荷驱动模拟[1]，会带来许多优点，如不受特定系统、特殊配置的局限，没有作业最长运行时间的限制等。

目前，评估计算网格时都是在并行计算机作业跟踪基础上直接构造网格工作负荷。文献[2]在康奈尔理论中心的作业跟踪基础上构造了 4 个工作负荷，每个工作负荷包含 10^4个作业，以轮转方式将工作负荷中记录的作业提交给不同节点。文献[3]从康奈尔理论中心的作业跟踪中截取 5 000 个连续的作业，分别作为 4 个节点的工作负荷，并使其中 2 个节点负载较重，方法是将作业运行时间伸长到原来的 1.7 倍。文献[4]从美国国家能源研究科学计算中心、圣迭戈超级计算中心等获取几个月的作业跟综，并附加作业的输入、输出数据量等信息，构造了几个工作负荷，分别提交到 12 个节点。这种直接构造的网格工作负荷存在若干不足，如没有考虑节点计算能力差别、没有考虑作业提交位置、直接（而不是由作业调度）改变节点负载等。通过工作负荷建模，可解决上述不足。

2 网格工作负荷模型

2.1 并行度和纯运行时间

运行中并行度可变的作业称作可模压作业；反之，称作刚性作业。可模压作业使并行度分布变得复杂，而刚性作业占并行作业的大多数，故这里只考虑刚性作业。

有些研究假设并行度服从均匀分布，这不符合实际。文献[5]在 3 个真实工作负荷基础上，用数据拟合法找出并行度的概率分布。作业分为串行和并行两类，并行作业的并行度又分为 2 的幂和非 2 的幂。假设作业为串行作业的概率为p_1，

作者简介：王庆江(1968－)，男，博士，主研方向：高性能计算，计算网格中的关键技术；张 琳，硕士

收稿日期：2006-02-23 **E-mail：**qjwang@ouc.edu.cn

为并行作业的概率为 1-p_1，并行作业中并行度为 2 的幂的概率为p_2×（1-p_1），非 2 的幂的概率为(1-p_2)×（1-p_1）。根据文献[5]的 4.2 节，令p_1=0.21，p_2=0.89。当并行度w为 2 的幂时，$\log_2 w$的概率累积分布函数是经过(0.8, 0)、(4.5, 0.86)和(7, 1)三点的两段直线，$\log_2 w$的概率分布如图 1。

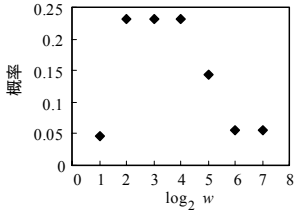


图 1 并行度w为 2 的幂时$\log_2 w$的分布

文献[5]中的工作负荷建模针对并行计算机，而本文的网格作业局限于一个节点运行，故可认为网格作业的并行度也服从图 1 所示的分布。节点A_k的并行计算能力可用一些测试软件（如NPB）获得，令为c_k，作业J_i在A_k上的运行时间为 t_{ik}^e，不妨假设 t_{ik}^e 与c_k成反比。 $c_k t_{ik}^e$ 称作J_i的纯运行时间，其含义是不依赖指派的“纯粹”的并行计算资源需求。根据节点之间并行计算能力的差别，由纯运行时间可获得作业在各节点上的运行时间，故只需对纯运行时间建模。

由文献[5]，并行计算机中作业运行时间符合超伽玛分布$p\Gamma(\alpha_1, \beta_1)+(1-p)\Gamma(\alpha_2, \beta_2)$。参考文献[5]表 3 的平均模型，令$(\alpha_1, \alpha_2, \beta_1, \beta_2, p)$ = (10.74, 37.96, 0.55, 0.37, 0.577)，对数转换后作业运行时间（$\ln l$）的分布如图 2。不妨假设图 2 所示的是网格作业的纯运行时间分布。

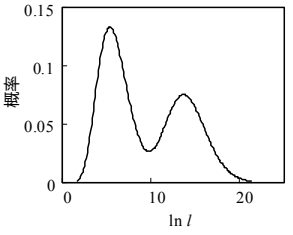


图 2 对数转换后作业运行时间的分布

2.2 提交位置和到达间隔

用户可将作业提交到任一个网格调度器。如果作业J_i提交到节点A_j上的网格调度器GS_j，则GS_j称作J_i的提交位置。

为模拟提交位置分布的不均匀，不妨设作业提交到各网格调度器的概率服从比率α的几何分布，这里网格调度器随机排序，且$\alpha \geq 1$。假设 8 个网格调度器的一个随机序列为(GS_6, GS_3, GS_7, GS_1, GS_2, GS_4, GS_5, GS_0)，若作业提交给GS_6的概率为q，则提交给GS_3的概率为q×α，提交给GS_7的概率为q×α^2，依次类推。$\alpha=1$时，提交位置服从均匀分布。α越大，分布越不均匀。

文献[5]对 3 个并行计算机的作业跟踪进行了分析，发现到达间隔服从伽玛分布$\Gamma(\alpha, \beta)$。根据文献[5]表 7，α=8.17，β=3.96，到达间隔的概率分布如图 3。一般可认为不同作业被提交到哪个网格调度器是彼此无关的，故各网格调度器收到作业的间隔服从彼此无关的伽玛分布，而网格的作业到达间隔分布是那些伽玛分布的叠加，即服从超伽玛分布。提交位置分布与各网格调度器上到达间隔的分布有关。当作业提交到某个网格调度器的概率较大时，该网格调度器上作业的间隔普遍较短。可直接假设网格的作业到达间隔分布如图 3，通过调整 α 控制各网格调度器上到达间隔的分布。

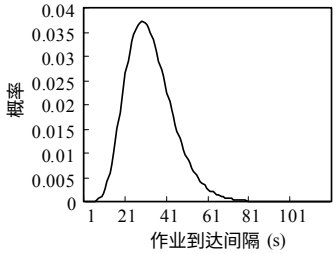


图 3 作业到达间隔的分布

文献[6]通过将作业到达间隔乘以一个因子来调整并行计算机的负载。类似地，为模拟网格负载的轻重，作业到达间隔可统一乘以因子 s，s 称作作业到达间隔的扩张因子。s 越大，网格负载越轻。

2.3 纯运行时间的估计因子

节点的许多本地调度（如装填法[6]）需要用户对作业运行时间进行估计。当对作业计算需求比较清楚时，用户对运行时间的估计会比较准确；反之，为确保作业顺利完成，需要对运行时间高估很多。以往，常假设估计值为真实值的若干倍[6]，这不符合实际。在不同节点上，作业的运行时间往往不同。通常，作业运行时间越长，估计误差越大，可假设相对误差为常值。运行时间与并行计算能力近似成反比，故只需对作业纯运行时间的估计误差建模。

假设作业J_i的纯运行时间为l_i，纯运行时间估计为l_i×$e(J_i)$，则$e(J_i)$称作J_i纯运行时间的估计因子。假设$e(J_i)$服从区间[E_1, E_2]上的均匀分布。E_2越大，l_i的高估程度可能越大；E_1<1 时，l_i可能被低估。调整E_1和E_2，就可调整作业纯运行时间的估计范围。

2.4 完成期限

作业完成期限是用户允许的最晚完成时刻。如果J_i的提交时刻为 t_i^a，纯运行时间估计为l_i×$e(J_i)$，指定的完成期限为 t_i^d，则通常 $t_i^d > t_i^a + l_i \times e(J_i)$。用户为$J_i$指定完成期限时主要依据$l_i$×$e(J_i)$，$t_i^d$ 可表示为 $t_i^a + B \times l_i \times e(J_i)$。在无法从真实网格中获得作业完成期限分布情况下，可假设B服从[B_1, B_2]上的均匀分布。

2.5 纯迁移开销

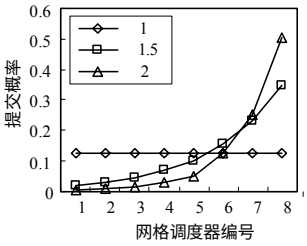
作业迁移开销就是在节点间传输作业可执行代码和待处理数据的时间，迁移开销对作业迁移起到了阻止的作用。

迁移开销依赖作业迁移时需要传输的数据量和网络传输能力。网络传输能力依赖带宽和延迟。当数据量很大时，可忽略延迟对迁移开销的影响；当数据量很小时，可忽略带宽对迁移开销的影响。用一组典型网络应用评估节点之间的网络性能，可把网络传输能力表示为一个综合指标。

迁移开销与网络传输能力综合指标近似成反比，只需对一种网络上的作业迁移开销（即纯迁移开销）建模。目前无法从真实网格中找到纯迁移开销的分布，不妨假设纯迁移开销服从[D_1, D_2]上的均匀分布。

3 应用实例

假设有 8 个网格调度器，α对提交位置分布的影响如图 4。$\alpha=1$时，提交到各网格调度器的概率相等；α 变大时，提交概率差别越来越大。可见，α 用于调整作业提交位置分布。调整扩张因子 s，连续 10 个作业的到达间隔变化如图 5。s 越小，作业到达间隔的均值将缩小，意味着网格负载变重。



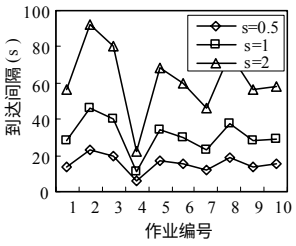
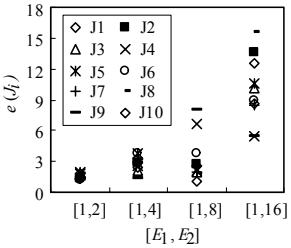


图 4 α对作业提交位置分布影响　图 5 扩张因子对作业到达间隔影响

调整E_1和E_2，连续 10 个作业的$e(J_i)$值如图 6。通常，纯运行时间不应低估，即$E_1 \geq 1$。E_2越大，$e(J_i)$值将分布在越大的区间上，即纯运行时间被高估更多。

令$e(J_i) \equiv 1$，$B_1=2$，$B_2=6$，连续 10 个作业的到达时刻（ta）、到达时刻+纯运行时间（tc）、到达时刻+完成期限（td）如图 7。这里tc代表作业的最早完成时刻，td代表用户指定的完成期限。这些作业的纯运行时间差别不大，但完成期限有较大差别，反映了完成期限或紧或松的情况。



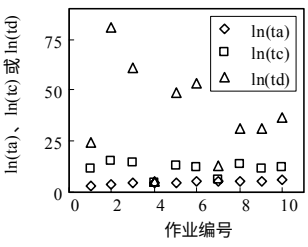


图 6 E_1和E_2对 10 个作业$e(J_i)$的影响　图 7 作业到达时刻、最早完成时刻和完成期限

给定不同的$[D_1, D_2]$，连续 10 个作业的纯迁移开销如图 8。D_1和D_2越小，作业的纯迁移开销越小；调整D_1和D_2，可以评估迁移开销对动态调度的影响。

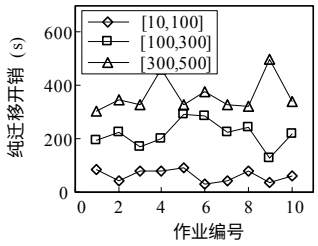


图 8 3 种区间下的纯迁移开销

4 结论

网格工作负荷是模拟环境下评估网格作业调度所必需的，然而目前无法从实际中获得有代表性的网格工作负荷，所以本文构建了网格工作负荷模型。

网格作业的属性除了包括并行度、运行时间、到达间隔外，还有提交位置、完成期限、迁移开销等。考虑到资源性能的差异，本文定义了不依赖作业指派的纯运行时间和纯迁移开销。由并行计算机的工作负荷模型，可获得并行度、纯运行时间和到达间隔的分布。另外，构建了作业提交位置、纯运行时间估计因子、完成期限和纯迁移开销的分布。应用实例展示了如何调整模型参数，以产生各种工作负荷，表明这里的工作负荷模型可用于网格作业调度的全面评估。

下一步，将研究工作负荷模型与实际工作负荷的吻合程度，并对模型进一步完善。

参考文献

1 Feitelson D G. Packing Schemes for Gang Scheduling[C]//Proc. of Workshop on Job Scheduling Strategies for Parallel Processing, 1996: 89-110.

2 Ernemann C, Hamscher V, Schwiegelshohn U. et al．On Advantages of Grid Computing for Parallel Job Scheduling[C]//Proc. of 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid, 2002: 31-38.

3 Subramani V, Kettimuthu R, Srinivasan S, et al. Distributed Job Scheduling on Computational Grids Using Multiple Simultaneous Requests[C]//Proc. of the 11th IEEE International Symposium on High Performance Distributed Computing, 2002: 359-366.

4 Shan Hongzhang, Leonid O, Rupak B. Job Superscheduler Architecture and Performance in Computational Grid Environments[C]//Proc. of ACM/IEEE SC´03 Conference, 2003: 44-58.

5 Lublin U, Feitelson D. The Workload on Parallel Supercomputers: Modeling the Characteristics of Rigid Jobs[EB/OL]. 2003-12-03. http://citeseer.nj.nec.com/lublin01workload.html.

6 Mu'alem A W, Feitelson D G. Utilization, Predictability, Workloads, and user Runtime Estimates in Scheduling the IBM SP2 with Backfilling[J]. IEEE Transactions on Parallel and Distributed Systems, 2001, 12(6): 529-543.

（上接第 69 页）

第 2 阶段为 PSM 到代码。本文主要工作是对于第 1 阶段以一对多双向导航关联为例，给出了一套从 UML 模型到 Java 模型的变换规则，并且用具体的实例证明了规则的正确性。

今后的工作将是：针对其他的关联模型（比如：一对一关联、单向导航关联等），定义其隐式和显式实现模式的变换规则；按照关联变换规则，实现一个具体的 UML 模型到 Java 模型变换工具；以及在模型自动转换的过程中，发现和消除不一致性。

参考文献

1 Laleau R, Polack F. Specification of Integrity-preserving Operations in Information Systems by Using a Formal UML-based Language[J]. Information and Software Technology, 2001, 43(12): 693-704.

2 Object Management Group. MDA® Specifications[EB/OL]. 2002-06. http://www.omg.org/mda/specs.htm.

3 Siegel J. Developing in OMG's Model-driven Architecture[EB/OL]. 2002-11. http://www.omg.org.

4 Rumbaugh J, Jacobson I, Booch G. UML参考手册[M]. 北京: 机械工业出版社, 2001.

5 王黎明, 柴玉梅. UML 中的关联关系及其实现模式[J]. 郑州大学学报(理学版), 2002, 34(3): 25-28.

6 Kleppe A, Warmer J, Bast W. 解析 MDA[M]. 鲍志云, 译. 北京: 人民邮电出版社, 2004.