计算机科学2003Vol. 30№. 7

分布式实时系统中任务调度问题的研究

胡金初
(上海师范大学计算机系　上海200234)

The Study of Schedule in Distributed Real-Time System

HU Jin-Chu
(Shanghai Normal University ,Shanghai 200234)

Abstract One way of increasing the computational speed is by using multiple computers operating together on a single problem. The overall problem is split into parts, each of which is performed by a separate processor in parallel. Writing programs for this form of computation is known as parallel programming. Equipped by electrical power monitoring system of Shanghai Hongqiao international airport, this report clarifies the methods of parallel processing.

Keywords Distributed system,Schedule,Algorithms,NP complete

1 引言

由于大规模集成电路的集成度越来越高,在一个系统中,要进一步大幅度提高单个计算机的性能就变得越来越困难。随着网络和通信技术的发展,现在使用一组计算机来完成一个分布式应用已经变得很普遍了。在实际应用中,采用多计算机的系统往往成为一种现实而又经济的选择。并行计算机和分布式系统的出现为高性能计算的应用开辟了广阔的天地。

但是,开发能够获取并行计算机性能的软件比在传统的单处理机上要困难得多,多处理机系统结构比单处理机复杂,决定了在组织计算上会出现比较多的问题。在多处理机系统中执行一个并行程序需要将任务分配到不同的处理机上,让它们并行地执行,从而获取较高的系统性能。如何将应用程序中的任务分配到处理机上,使并行程序的执行时间最小化,就成了一个重要的问题。任务到处理机上的分配称为映射,找到能够最小化执行时间的映射通常被称为任务到处理机的映射问题,这个问题是并行计算中最重要的问题之一。

2 调度问题

任务到处理机的映射是要以一种有效的方式将任务分配到处理机上,通过在处理机上的每个任务的运行,达到整个程序执行时间的最小化。如果处理机的数量多于任务的数量,那么分配是很容易进行的。但通常情况是任务的个数要多于处理机的个数。在这种情况下,为了充分利用所有的处理机,任务应该均匀地分布在各台处理机上,以获得最大程度的并行性。同时随着任务在并行机上的分布,任务间的通信也相应增加了。分配策略就需要考虑将具有较大通信量的任务分配到同一个处理机或邻近的处理机上。并行性和通信问题有时候是矛盾的,因为分配策略的质量是由系统的效率决定的,它和计算量与通信量都有关系。任务到处理机的映射问题在理论上可以用带权的有向图(DAG)来描述,图中的顶点表示程序中的任务,图中的边表示任务间的通信,边上的权值表示通信的费用。

当给定了任务图和处理机图后,映射的目标就是要完成一个分配:将 m 个任务映射到 n 个处理机上,简单的枚举分配法会产生 m^n 个状态空间。要获得最优解需要搜索巨大的状态空间,这在计算机上是难以处理的。过去的大多数算法都是次最优的,这是因为最优映射是一个 NP 完全的问题,时间复杂性很高,目前尚无有效的算法。

在一个多计算机系统中,根据某个策略来分配一组任务,使其获得最好的结果,这就是调度问题。这个问题在不同的领域中可以由许多不同的方法来进行描述。根据调度策略所要达到的目标,调度技术可被分为局部和全局的。局部调度被用在并发任务到单个处理机时间片的调度上,操作系统通常处理局部调度。大多数研究人员关心的是全局调度,全局调度要找到一个最好的可能方法来对给定的任务负载进行组织,以使得处理机的执行时间最短。

任务调度可以有很多方式进行分类。一种分类是分成静态调度和动态调度。采用静态调度需要对系统的所有信息都已知,例如具有约束关系的任务图和数据交换产生的额外开销等。静态调度方法的目标是最小化并行程序的总执行时间,但如果在程序执行前不知道任务图,那么程序就不能满足要求。例如,条件分支是可能引起不确性的一个程序结构,只有在程序的运行过程当中才能确定分支的走向,使用静态策略是无法解决这个问题的,这时就要利用动态调度了。当在程序执行前无法获得程序的全部信息时,并行系统就要动态地对任务进行调度。动态调度的缺点是这种方法很难达到全局的最优化,并且在程序运行的过程中进行调度会产生一些额外开销,降低了系统的效率。

在调度方式上又可以分成"优先调度"和"非优先调度"两种。所谓"优先调度"是指把某个任务分配给处理机后,允许其他优先级高的任务中断这个任务的执行。相反,在"非优先调度"中,当前执行的任务可以不中断地一直运行到任务结束。

作业调度和任务调度也是两种不同的调度类型。作业调度是指包含很多任务的作业被分配到一个处理机上,而任务调度是指不同任务被分配到并行处理系统的不同处理机上。

人们在很多并行系统上对任务调度作了研究,成了应用程序处理的一个重要内容。在这样的环境中,调度器将每个程序模块分配到一个处理机上,以便在处理机利用率和吞吐量上获得所需的性能。在调度问题上可以分成3个层次:高层调度,也称为作业调度,从所有对系统提出处理要求的作业中选

择合适的作业分配给各个处理机；中层调度，根据系统对负载的响应情况来决定临时挂起或者激活任务，使系统运行更平稳；低层调度，在某个时间段内决定分配给处理机下一个就绪的任务。常用的调度策略有：先进先出、轮转法、最短作业优先和最短剩余时间等，根据系统的需要，选择不同的调度策略。

在异构并行系统中，由于处理机性能的差异会引起负载的不平衡，从而导致系统性能的下降，因此在异构的系统中，除了通常的调度方法之外，还需要在系统层次上有一个调度层，用来监控系统的通信瓶颈等情况，了解所有任务的执行情况，维护系统范围内的负载平衡。由于硬件结构的异构性，使得系统内的排队问题更明显了，这给调度策略带来了不利的限制因素。在调度时，还需要知道任务的类型以及适用的机型，因此与调度相关的问题在异构的计算机上就变得更复杂了，这还有待人们进一步的研究。

并行程序的有效执行在很大程度上取决于程序到模块的有效划分以及调度这些模块在一组处理机上的执行。影响程序并行性的因素有很多，其中主要的因素包括程序的划分、计算负载在各处理机上的平衡和数据通信所产生的额外开销等。对于给定的一个并行程序，程序的划分是将程序分成若干个单元，这些单元可以分配给处理机并行地执行，以使得程序的全部计算时间是最小的，其中包括处理机间的通信开销。

3 调度实例

电力系统中大量应用计算机的监控技术，以提高电力供用电系统的性能，并能对出现的问题及时作出反应。在供用电中产生的大量电力数据要及时采集，及时处理，并对相关供电设备进行快速精确的控制。例如，某一路电量的计算可以通过电功率的积分获得：

$$E_i = \int P_i(t)dt$$

因此在为电力数据监测控制配备计算机设备时，精确、可靠和实时性就成了基本的要求。我们在对虹桥国际机场电力监控和信息管理系统的研制中，为了满足用户的高要求，采用多计算机技术，组成分布式处理，保证了系统的高性能和安全可靠的运行。

该系统有南北两个变电站，各有两路35kV 的进线，并分别提供40和48个分路的输出，是一个大型的用电单位。由3台计算机组成一个分布式的多处理机系统，在供用电过程中产生的大量电力数据要及时送到各个计算机上去处理。这些数据可以分成采集、处理、分析、打印和输出等几个步骤。分析时，假定在一个系统中有 n 个作业 $R=\{R_1,\cdots,R_n\}$ 要处理，而每一个作业 R_i 又可以分成 m 个任务 $R_i=\{R_{i1},\cdots,R_{im}\}(1\leqslant i\leqslant n)$，计算机系统本身由 m 台计算机 $P=\{P_1,\cdots,P_m\}$ 组成，任务 R_{ij} 由处理机 P_j 来处理，P_j 完成任务 R_{ij} 需时间 t_{ij}。n 个作业的调度问题是为任务 R_{ij} 在处理 P_j 上安排工作时间表，用最少的时间来完成这一批作业。这里的条件是：

1. 任务 R_{ij} 在处理机 P_j 上执行；

2. 对任一任务 R_{ij}，当 R_{ij} 没有完成时，$R_{i,j+1}$ 就不能执行；

3. 在同一时间内，没有哪一台处理机可以接受两个或者两个以上的任务；

4. 任务 R_{ij} 对应的处理时间为 t_{ij}。

对于虹桥国际机场电力监控系统有3台计算机，在某一时刻有4个作业要处理。各作业所需的不同处理机的时间由矩阵 T 给出：

$$T=(t_{ij})_{4\times3}=\begin{bmatrix}8&2&3\\4&7&1\\3&6&10\\5&9&5\end{bmatrix}$$

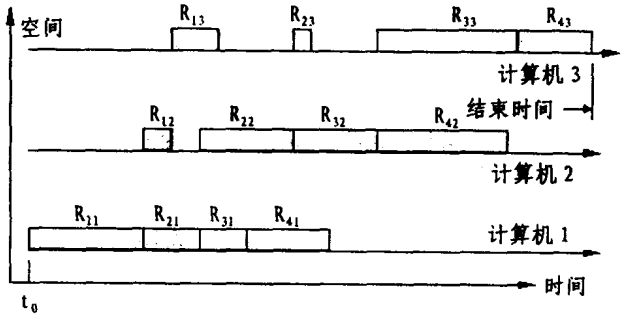

图1 一种调度方案

根据图1的调度方案，可以很容易计算出完成这批任务的处理所需要的时间为：

$t=8+4+13+15=40$

即总共需要40个单位时间。从图中我们不难看出，3号计算机有一段时间是处于等待状态，从而拉长了任务完成的时间，在分配上不是很合理。我们的目的是要找到最佳的调度方案，使完成任务需要的时间最短。我们采用下面的方法来解决。

每一个任务在开始执行时，可以估计出它在理想情况下所需要的最少执行时间：

$$t_i = t_{i1}+\sum_{j=1}^{4}t_{j2}+\min\{t_{j3}\}\mid\forall\ j\neq i$$

式中表示，当某一个作业 R_i 在计算机1上完成了第一个任务以后，计算机2就可以开始工作。理想情况是计算机2一直不空闲，把所有要在它上面处理的任务全部完成。最后，在计算机3上再处理的任务是 R_{j3}，它对应的时间 t_{j3} 同时满足以下条件：

$$t_{j3}=\min\{t_{j3}\}\mid\forall\ j\neq i$$

处理时间受到的另一个约束条件是：3号处理机在尽可能早的时刻开始处理任务，然后马不停蹄地完成全部任务，这是它所需要的最少时间。即

$$t_i=\max\{t_{i1}+\sum_{j=1}^{4}t_{j2}+\min\{t_{j3}\}\mid\forall\ j\neq i, t_{i1}+t_{i2}+\sum_{k=1}^{4}t_{k3}\}$$

按照上面的方法，可以计算出各作业处理时，在理想情况下最少需要的时间：

$$R_1: t_1=\max\{t_{11}+\sum_{j=1}^{4}t_{j2}+\min\{t_{j3}\}\mid\forall\ j\neq 1, t_{11}+t_{12}+\sum_{k=1}^{4}t_{k3}\}=\max\{8+24+1, 8+2+19\}=\max\{33,29\}=33$$

同理可以求出：

$R_2: t_2=\max\{31,30\}=31$

$R_3: t_3=\max\{28,28\}=28$

$R_4: t_4=\max\{30,33\}=33$

从上面的分析可以知道，在计算机1上先处理第3个作业比较好，有可能获得最佳的分配策略，花费最少的处理时间。然后排除1号作业，在剩余的作业中选择第2个要处理的任务。经过进一步的分析，可以得到图2。在实际使用中，一般采用分支定界法对这一类问题进行分析和处理。

从图2中的分析可以知道，采用 $R_3\rightarrow R_4\rightarrow R_1\rightarrow R_2$ 的处理

（下转第146页）

人的各种智能表现在信息加工的心理功能上都以记忆与学习为核心，在生理基础上都可还原为神经元对信号进行加工的物理与化学过程，及其连通网络从宏观上的整合。因此我们认为人工智能类脑模型的核心是基于神经元信息处理机制的集群表示模型及其系统的动力学行为。所以我们将围绕记忆与学习开展研究，并逐步拓展。

我们需要将宏观模型和模型的实现基础作一个区别，上层模型(整体)必然更靠近心理学模型，下层实现(细节)必然更靠近生理学模型，它们间的过渡与衔接是一个心理生理学的示例。记忆机制是融合表示问题的关键，我们先考察记忆与神经元间信息传递的生理机制；再设计基于神经元学说的记忆表达模型，包括网络的结构模型、模型的数学描述、动力学过程的数学模型等；再设计基于结构特征的记忆编码算法；然后进行包括稳定性、分辨率等指标在内的网络性能分析；最后以得到的模型为核心，逐步建立对人工智能各类问题的解释。

参 考 文 献

1 Kirsh D. Foundations of AI: the big issues. Artificial Intelligence, 1991,47:4～10

2 危辉. 认知相关性与智能模型构造的系统观点:[博士后研究工作报告]. 杭州:浙江大学研究生院,2000

3 Taddei-Ferretti C, Musio C. Downward processes in the perception representation mechanisms. Singapore: World Scientific Publishing, 1998

4 Koch C. Biophysics of computation: information processing in single neurons. New York: Oxford University Press, 1999

5 Burdet G, Combe P, Parodi O. Neuronal information processing: from biological data to modeling and applications. Singapore: World Scientific Publishing, 1999

6 Riegler A, Peschl M, von Stein A. Understanding representation in the cognitive sciences: Dose representation need reality?, New York: Kluwer Academic / Plenum Publishers, 1999

7 Uttal W R, et al. Computational modeling of vision: the role of combination. New York: Marcel Dekker Inc., 1999

8 Squire L R, Kosslyn S M. Findings and current opinion in cognition neuroscience. Cambridge, Massachusetts, The MIT Press, 1998

9 Gordon L. Fain, Molecular and cellular physiology of neurons. Cambridge, Massachusetts, Harvard University Press, 1999

10 Bower J M. Computational neuroscience: trends in research. New York: Plenum Press, 1997

11 Bower J M. Computational neuroscience: trends in research. New York: Plenum Press, 1998

12 Quinlan P T. Structural change and development in real and artificial neural networks. Neural Networks, 1998, 11:577～599

13 Haken H. Synergetic computers and cognition: a top-down approach to neural nets. Heidelberg, Springer-Verlag, 1991

14 Levitan I B, Kaczmarek L K. The neuron cell and molecular biology(2nd Edition). New York: Oxford University Press, 1997

15 Longstaff A. Instant notes in neuroscience. London: BIOS Scientific Publishers Limited, 2000

16 危辉, 何新贵. 表象式直接知识表示. 计算机学报, 2001, 24(8):891～896

17 危辉, 潘云鹤. 从知识表示到表示:人工智能认识论上的进步. 计算机研究与发展, 2000, 37(7):819～825

18 危辉. 一种基于结构学习和迭代自映射的自联想记忆模型. 软件学报, 2002, 13(3):438～446

(上接第131页)

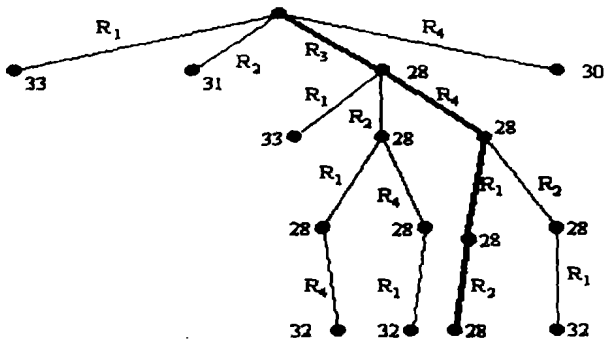

图2 分支定界分析

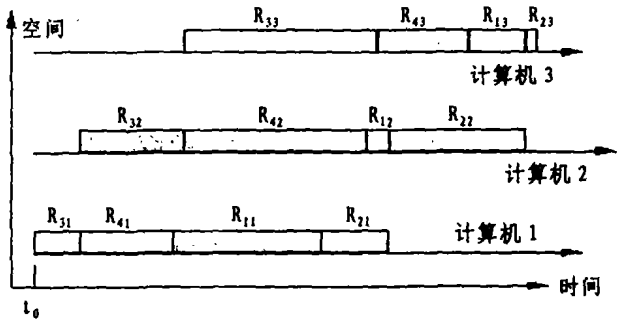

图3 另一种调度方案

顺序，可以获得最佳的效果。把处理时间从原来的40个单位时间减少到28个单位时间。在系统中实际运行的情况见图3，可见提高了系统的并行性。

研究网络并行处理的体系结构和实现技术，是今后计算机应用的一个重要方向，包括多线程的体系结构和预取、后存技术，在网络计算环境下算法的高效实现问题。

参 考 文 献

1 Pritchard J. COM and CORBA Side by Side Architecures, Strategies, and Implementations Addison Wesley Longman. Inc. 1999

2 Wu Jie. Distributed System Design. CRC Press LLC, 1999

3 Buyya R. High Performance Cluster Computing, Volume1and 2, Prentice-hall, Inc. 1999

4 胡金初. 分布式处理在实时应用系统中的实现. 计算机工程, 2002, 12