并行系统性能影响因素分析*)

Analysis of Performance Affecting Factors of Parallel System

王与力　杨晓东

(国防科技大学计算机学院　长沙410073)

Abstract　In this paper, we discussed the definition method of the components of the execution time model of parallel system, and the relationship of the time components with the system characters, we proposed the concept of critical path and analyzed the effects of the characters of program, machine and runtime environment on the time of computation, communication, shared memoty access, I/O and synchronization on the critical path, so as to expose the relationshepe of the performance and system character and furthur the association of performance model and system model, thus facilitate the analysis and improvement of parallel system performance.

Keywords　Parallel system, System character, Performance character, System model, Performance model

1. 引言

研制并行系统的根本目的在于克服单个处理机的性能限制，满足应用问题特别是所谓挑战性问题中的诸多性能需求。由于性能在并行系统中的重要性，以评价性能特征，确定性能瓶颈并提出性能改进措施为目的的性能研究在并行系统的设计和使用中也起着重要指导作用。并行系统性能研究常用的三种基本方法是测试方法、模拟方法和分析方法。由于测试方法只能适用于已经设计好的并行系统，而模拟方法在适用于大规模的并行系统时，模拟器的构造十分复杂，模拟器的运行也需要极大的时间和空间代价，因此以建立和分析数学模型为主要手段的分析方法被认为是能适用于并行系统的早期设计阶段，并且模型复杂性和模型分析的时空代价都较小的有效的性能研究方法。分析方法最本质的特征是抽象，即抽象地用一个或一组数学等式来反映系统的性能特征。高度抽象的数学模型能够更好地反映一般意义上的并行系统的性能特征，但对一实际的具体的并行系统而言，数学等式中的各参数如何定义，如何由实际的并行系统中产生，如何与具体的系统特征建立联系，往往很不明确。因此，如何为一实际的并行系统建立分析性性能模型，如何利用性能模型指导实际的并行系统的性能分析和改进是利用分析方法进行并行系统性能研究时面临的难题。本文对影响性能的系统特征进行了分析，从而建立起性能特征与系统特征的联系并由此建立性能模型与系统模型之间的联系，即：

性能模型↔性能特征↔系统特征↔系统模型

以便在对一实际的并行系统进行分析性性能研究时，首先为实际并行系统建立能够反映有关系统特征的系统模型，由系统模型建立反映有关性能特征的性能模型，然后，由性能模型分析影响性能特征的实际系统的系统特征，从而定位性能问题，提出改进指导。

本文分析了并行系统的性能特征，提出了用关键路径概念定义时间模型中的时间分量的思想，分析了并行系统的性能特征与系统特征之间的联系，提出并行系统的系统模型和性能预测方法的初步设想，最后部分对本文进行了总结。

2. 性能特征分析

并行系统的性能是指并行系统在时间和速度方面的特性，并行系统性能指标是用来直接或间接地刻画系统性能的有关度量，性能研究的根本点是用一组性能指标来反映并行系统有关方面的性能特征。由于并行系统的层次性，性能研究中要涉及系统的不同层次、不同方面的性能指标。总的来说，性能指标可划分为系统级、程序级、机器级、运行环境级四个方面。

系统级的性能指标包括加速比、效率、可扩性等。其中加速比是指程序的串行执行时间和并行执行时间的比值；效率为系统的加速比与系统中处理机数目的

*)本课题得到863高技术基金资助。王与力　博士研究生，主要研究方向为并行系统性能评价。杨晓东　教授，博士生导师，主要研究领域为高性能计算机体系结构，互联网络，性能评价、可靠性理论等。

比值；可扩性则是反映系统的问题规模和机器规模变化时，系统性能的变化特性。

程序级的性能指标包括程序结构的不同部分的执行时间、速度及其相互关系，如并行程序及其各任务的执行时间，任务中各代码段的执行时间、以及并行程序中有关部分的负载均衡程度（指相互间的负载规模的差异程度）。程序级的性能指标还包括程序行为的不同方面的执行时间，速度及其相互关系、如计算时间（速度），通信时间（速度），访存时间（速度），同步时间（速度），I/O时间（速度），以及计算通信比一计算时间（量）与通信时间（量）的比值，计算访存比一计算时间（量）与访存时间（量）的比值等。

机器级的性能指标包括用来反映机器的不同功能部分的速度的度量，如计算速度、各级访存速度、各部分通信速度（接口，网络）、同步速度，I/O速度等。速度方面的度量根据其求值方式的不同可分为三类，即：理论上的峰值速度、benchmark 测试速度、应用程序实际速度。机器级的性能指标还包括用来反映机器的效率和利用率等方面的度量。效率是指系统有关实际速度或 benchmark 速度与峰值速度的比值，如计算效率、通信效率等。利用率是指系统有关部分的实际利用时间与系统总的活动时间的比值，如 CPU 利用率、网络通道利用率等。

运行环境级的性能指标包括有关的库操作开销、运行时的管理操作开销以及运行时的重叠、拥塞或等待时间，如各类通信操作开销、cache 一致性管理开销、锁等待时间、网络通道等待时间等。

由并行系统各级层次的性能指标类别及其定义知，对并行系统的性能而言，时间是关键，各级性能指标中的大多数是直接反映某一方面的时间特征的，而其它的性能指标几乎都是有关方面的时间函数。

由于时间在并行系统性能特征中的重要性，下面我们将考虑以程序并行执行时间模型为主的并行系统的性能模型。

3. 时间模型与关键路径

程序的并行执行时间（T_P）是指并行程序中从第一个任务开始执行到最后一个任务结束时所经历的时间。分析方法中研究时间的常用方法是将时间划分为若干分量[1~3]。一种方式[2]将 T_P 划分为串行部分执行时间和可并行部分执行时间；另一种常见方式将 T_P 划分为计算时间和并行开销时间，文[1]中将 T_P 划分为串行部分时间、可并行部分时间、通信时间和同步时间。不失一般性，本文将 T_P 划分为计算时间 Tcomp（包括 CPU 中的运算时间和私有数据访问时间），通信时间 Tcomm，共享数据访问时间 Tsmem，同步时间 Tsync，I/O 时间 $T_{I/O}$，即：

$$T_P = T_{comp} + T_{comm} + T_{syn} + T_{I/O} + T_{smem} \quad (1)$$

时间划分的好处在于能直观地反映出系统中不同类型活动（如计算、通信、同步、I/O、共享访问等）所占用的时间，但存在两方面问题，其一为：时间划分中各分量如何定义，或者说各时间分量如何与实际系统的并行程序中有关任务或任务中的有关部分联系起来；其二为：各时间量受哪些系统特征的影响，如果解决了这两个问题，则由时间划分不仅可知系统中各方面活动的占用时间，而且可将各方面时间与有关任务及其中代码段联系起来，并且还可反映出有关的系统特征对各方面活动时间的影响，这对系统的性能特征分析和性能改进指导都极具价值。

对第一个问题，我们提出的一种解决方法为：假设并行程序的工作由 Nt 个任务（task）来完成，每一任务的工作由 N_P 个阶段顺序执行完成，任务 i 的执行路径表示为：$Et_i = \langle E_{P1}, E_{P2}, \cdots, E_{PNP} \rangle$，$EP_j = \langle$Pid，Ptype，Ltime，Pre$\rangle$描述任务 i 中第 j 个阶段的执行信息，其中，Pid 为阶段的序号，Ptype 为阶段的类型（包括计算，通信，同步，I/O，共享访问等类型），Ltime 为阶段的执行时间，Pre 为阶段的依赖关系（包括通信、同步、数据共享等关系），整个并行程序的执行全过程表示为序列：$E = \langle E_1, E_2, \cdots, E_n \rangle$，$E_i = \langle$Tid，Pid，Ptype，time$\rangle$，Tid 为任务序号，Pid 为阶段序号，Ptype 为阶段类型，time 为阶段持续时间。E_1. Pid 为并行程序并行执行中，最先开始执行的阶段，En. Pid 为并行程序并行执行中，最后完成的阶段，序列中的各阶段在时间上呈串行关系。称序列 E 为并行程序执行的关键路径，由关键路径 E，式(1)中的各分量可定义为：

$$T_x = \sum_{E_i.P_{type}=x} E_i.time, i = 1, \cdots, n \quad (2)$$

即式(1)中各分量分别为并行程序执行的关键路径上对应活动的执行时间之和。由式(2)，可将各分量与并行程序中有关任务的有关阶段的对应代码段联系起来，

此解决方法的关键是如何产生任务执行路径 E_e 和程序执行关键路径 E。并行程序中，各任务都是串行执行的，因此任务的执行路径可由分析任务代码并求取各阶段时间而获得。并行程序各任务间的并行和各种交互形成的复杂关系，使得关键路径的求取比较复杂，其具体求解方法我们将另文讨论。

4. 系统特征分析

本节分析时间模型式(1)中各时间分量受并行系统中哪些系统特征的影响，以期建立性能特征（各时间分量）与系统特征之间的联系。

式(1)中的计算时间分量可用等式表示为：Tcomp = Acomp/Bcomp。其中 Acomp 为等效计算量，Bcomp 为等效计算速度。这里引入等效概念是因为程序中常常涉及不同类型的计算（如整型，浮点，双精度等等），而机器中不同的指令类型常常需要不同的执行拍数

(即不同的执行速度),当仅用一个量描述计算量或计算速度时,需要在不同类型之间作等效变换。这里仅讨论系统特征对时间分量的影响,故对具体的等效变换方式不作考虑,由于 Tcomp 为关键路径上的计算时间,因此 Acomp 也应为关键路径上的计算量,这一计算量首先与程序中总的计算量 ATcomp 和程序中计算的并行度 Pcomp 有关,可粗略表示为:Acomp = ATcomp/Pcomp。该式的潜在假设为:各任务中的计算负载完全均衡且调度算法为一台处理机上最多分配一个任务,在负载不均衡(不均衡度为 f)和调度不均衡(最多往一台处理机上分配 s 个任务),则 Acomp 的稍准确的描述应为:Acomp = ATcomp/Pcomp * f * s。Bcomp 与处理机的 cpu 频率 Ccpu 及其指令并行性 Pinstr 有关,可粗略表示为:Bcomp = Ccpu * Pinstr,该式的潜在假设为任务中能提供足够的指令并行,且无因相关和访存等引起的 cpu 停顿。任务中的实际指令并行性与其中的计算(指令)类型及各类指令的比例以及指令中的数据访问模式有关。访存停顿时间与任务中的数据量,数据访问模式和机器中的存储结构,各级存储的访问速度以及运行时存储结构中的数据状态有关。因此,影响计算时间分量 Tcomp 的系统特征可归纳为:

程序特征:计算类型,计算量,并行度,负载均衡性,数据量,局部数据访问模式

机器特征:计算速度(cpu 频率,cpu 指令并行度,指令 CPI),访存速度,存储结构

运行环境:任务调度规则,数据状态,存储状态

式(1)中的通信时间分量 Tcomm 也可粗略地用分析性等式表示为:Tcomm = Acomm/Bcomm。其中 Acomm 为等效通信量,Bcomm 为等效通信速度。Tcomm 为关键路径上的通信时间,故 Acomm 也应为关键路径上的通信量。Acomm 由程序中总的通信次数和每次通信的数据量(C_i),以及任务间通信的并行度(Pcomm),机器中的通信并行度(Mcomm),以及计算和通信的重叠度 OL 有关,可直观地表示为:$Acomm = (\sum C_i)/\min(Pcomm, Mcomm) * OL$。通信速度 Bcomm 与通信操作类型,通信操作的实现方式以及机器的通信能力:包括处理机的运算速度,接口通信能力(接口带宽和延迟),网络通信能力(路由器和通道的带宽和延迟,源与目标间的距离)有关,Bcomm 还与通信状态包括网络状态和任务状态有关,网络状态指网络中路由器和通道的忙闲,它与程序中的通信模式,机器中的网络拓扑,以及网络中的路由和切换方式,通信操作的实现方式有关。程序中的通信模式指各通信操作间的时序关系(串行,并行),以及通信操作的源和目标分布。因此,影响通信时间分量 Tcomm 的系统特征可归纳为:

程序特征:通信量,通信次数,通信类型,通信次序,通信伙伴(源与目标)

机器特征:处理机速度,接口通信性能(接口延迟与带宽),网络通信性能(延时,带宽,距离),通信子系统结构,通信操作方式

运行环境特征:通信规则(路由,切换方法),通信库操作算法,网络状态,任务状态,通信库操作开销

共享数据访问时间分量 Tsmem 可简要地表示为:Tsmem = Asmem/Bsmem,其中 Asmem 为关键路径上的等效共享数据访问量,Bsmem 为等效共享数据访问速度。共享数据访问量 Asmem 与程序中共享数据总量和共享数据在各任务间的分布以及共享数据访问次数有关。共享数据访问速度 Bsmem 与机器中基本的访存和通信能力有关,而且受机器的系统结构(存储结构,网络结构),共享访问类型以及基本共享访问操作的实现方式影响。影响共享访问速度的重要因素还包括:共享访问模式(访问目标分布,共享伙伴,访问时序),共享数据状态(共享数据分布,共享数据副本数,数据存储状态),Cache 一致性协议操作及其开销。因此,影响共享访问时间分量的系统特征可归纳为:

程序特征:共享数据量,共享数据分布,共享访问次数,共享访问类型,共享访问时序,共享访问数据对象,共享访问伙伴

机器特征:计算能力,通信能力,存储能力,存储结构,网络结构,共享访问操作实现方式

运行环境特征:Cache 一致性机制,数据状态,存储状态

同步时间分量 Tsync 可表示为:Tsync = Async * Bsync,Async 表示关键路径上的等效同步次数,Bsync 表示等效同步开销。Async 与程序中的同步次数和参与同步的任务数目,各同步操作间的次序与并行性有关,Bsync 与机器中的基本同步操作的实现方式及其性能有关。此外,Bsync 还与同步操作类型,同步操作涉及的处理机数目,参与同步的各任务的运行状态以及同步库操作的实现方式及开销有关。因此,影响同步时间分量的系统特征可归纳为:

程序特征:同步操作类型,次数,次序,伙伴

机器特征:同步子系统结构,基本的同步操作方式及其性能

运行环境特征:同步库操作算法,同步开销,任务状态,同步子系统状态

I/O 时间分量 $T_{I/O}$ 可分析性地表示为:$T_{I/O} = A_{I/O}/B_{I/O}$,其中 $A_{I/O}$ 为关键路径上的等效 I/O 量,$B_{I/O}$ 为等效 I/O 速度。$A_{I/O}$ 与程序中的 I/O 次数,I/O 数据量,各任务中 I/O 操作的并行性和均衡性有关。$B_{I/O}$,则与机器中 I/O 子系统的结构,基本 I/O 操作的实现方式及性能,程序中 I/O 操作的类型,高层 I/O 操作的实现方式与开销,I/O 操作的源和目标分布,I/O 子系统状态以及参与 I/O 的各任务的状态有关。因此,

与I/O时间分量有关的系统特征可归纳为：

程序特征：I/O数据量，I/O操作次数，I/O操作类型，I/O操作时序，I/O操作的并行性与均衡性

机器特征：I/O子系统结构，基本I/O操作方式，基本I/O操作性能

运行环境特征：I/O子系统状态，任务状态，高层I/O操作实现方式与开销

由以上分析可知，影响并行程序执行时间的系统特征来自并行程序，并行机器和运行时环境三个方面，其中并行程序特征包括需求和结构两方面，需求指程序中计算、通信、共享访问、I/O，同步各类操作的量（计算量，数据量）、次数、类型，结构指各方面操作的并行性、均衡性、及操作的次序。并行机器方面特征包括机器的计算，存储，通信，I/O与同步子系统结构，各基本操作实现方式及性能。影响程序执行时间的运行环境因素包括状态（处理机状态、任务状态、数据状态、Cache状态、网络状态、I/O结点状态），规则（OS或运行库中通信、同步、I/O、共享访问操作算法、任务调度算法、Cache一致性协议、网络路由和切换方式）以及各有关操作开销。

在对并行系统进行预测性性能研究时，我们可以根据性能特征与系统特征的联系建立以程序需求与程序结构为内容的并行程序模型，以机器结构、基本操作实现方式、基本性能数据为内容的并行机器模型，以有关的运行时状态、运行时管理规则、运行时库操作实现方式及有关开销为内容的运行环境模型。对由程序模型，机器模型和运行环境模型构成的并行系统模型进行以关键路径分析方法为主的性能求解，从而在不要求系统实际运行，不需要进行复杂的模拟的情况下，对包含程序、机器，运行环境在内的完整的并行系统进行全面的性能预测和性能改进。（详细内容将另文介绍）。

结束语 本文分析了并行系统的层次性的性能指标体系，并针对其中最根本的性能指标—执行时间，讨论了分析性时间模型中各时间分量的实际意义，定值方法以及影响时间分量的系统特征，提出了将并行程序执行时间与并行程序的代码和执行活动相联系的关键路径概念，建立了并行系统中并行程序、并行机器以及运行时环境三方面的系统特征与各时间分量之间的联系，初步提出了建立包括并行程序模型、并行机器模型、运行时环境模型在内的并行系统模型以及通过对系统模型进行关键路径分析来预测并行系统的性能，指导系统性能改进的设想，在下一步的工作中，我们将致力于该方法的深入分析与完善。

参考文献

1 杨晓东. MPP系统的粒度匹配加速比模型. 计算机学报，1997(10月增刊)

2 Huang Kai. Advanced computer architecture: parallelism, scalability, programability New York McGraw Hill 1993

3 Xiao Xiaoqiang, Jin Shiyao, et al. The effect of communication performance on the speedup of MPP. In: Proceeding of HPC Asia 2000. May 2000. 399～402

（上接第57页）

上述算法的主要操作是比较两图的边，最坏情况下需进行$(n^2-n)/2$次，故该算法的最坏时间复杂度为$O(n^3)$，n为图中的节点数。2D-strings匹配的时间复杂度为$O(M)+O(N^2*lp^3)$，M,N表示匹配表中的项数$(M,N>n)$，lp为匹配表的最大长度。一些新提出的匹配算法最坏时间复杂度也是NP的[7]。所以，上述算法不仅简单，而且比所有已知的2D-string匹配算法具有更高的效率。

小结 本文描述了表示两对象间空间关系的四种不同方法。所提出的精化的二维投影间隔关系模型在二维投影间隔关系基础上加入了δ_m，χ_m和φ_m，使得系统可以回答诸如nearby，faraway等语义空间关系，比前者具有更高的精确度。但该空间关系由于加入了额外的参量，因此也额外增加了计算费用，在实际应用中，需要权衡精确度与效率，以选择合适的空间关系模型。

空间关系图是模型化图像对象间空间关系的有效方法，本文描述了该图的简单构造方法，并描述了支持精确匹配和相似性匹配的检索算法，该算法比所有已知2D-string串匹配算法具有更高的效率。

参考文献

1 Peuquet D J, Ci-Xiang Z. An algorithm to determine the directional relationship between arbitrary-shaped pologons in the plane. Pattern Recogination, 1987, 20(1): 65～74

2 Takahashi T, Shima N, Kishino F. An Image retrieval method using in queries on spatial relationships. Journal of Information Processing, 1992, 15(3): 441～449

3 Chu W W, Cardenas A F, Taira R K. KMeD: A Knowledge-Based Multimedia Medical Distributed Database System. Information Systems, 1995, 20(2): 75～96

4 Hsu C-C, Chu W w, Taira R K. A Knowledge-Based Approach for Retrieving Images by Content. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(4): 522～532

5 Egenhofer M J, Franzosa R D. Point-set topological spatial relations Int. J. Geographical Information Systems, 1991, 5(2): 161～174

6 Aslandogan Y A, et al. Design, Implementation and Evaluation of SCORE (a System for Content based Retrieval of Pictures). IEEE Computer, 1995, 28(2): 280～287

7 Freksal C. Temporal Reasoning Based on Semi-Intervals. Artificial Intelligence, 1992, 54(1-2): 199～227

8 Nabil M, Shepherd J, Ngu A H H. 2D-Projection Interval Relationships: A Symbolic Representation of Spatial Relationships, Advances in Spatial Databases: Fourth Int'l Symp., SSD'95, Lecture Notes in Computer Science, No. 951, Springer-Verlag, 1995. 292～309