

动态可扩展机群系统性能分析

吴潜蛟¹, 潘若禹², 武奇生³

(1 长安大学 信息工程学院, 陕西 西安 710061; 2 西安邮电学院 通信工程系, 陕西 西安 710061;

3 长安大学 电控学院, 陕西 西安 710061)

摘 要: 局域网机群系统是建立高性能计算环境的有效途径. 定义机群系统抽象结构及性能模型是实现上述目标的关键. 给出了机群 m/w 计算模型系统性能分析模型, 通过大规模 Monte Carlo 仿真应用实验, 证性能模型的有效性, 研究成果对于机群系统建立提供有效的组织、规划、实现方法.

关键词: 动态可扩展机群系统; 主从(m/w)计算模型; SPMD 应用; 并行计算任务粒度

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2010)10-0083-04

Performance Analyzing Model for Scalable Cluster

WU Qian-jiao¹, PAN Ruo-yu², WU Qi-sheng³

(1 School of Information Engineering, Chang'an University, Xi'an 710061, China;

2 Department of Communication Engineering, Xi'an College of Post & Telecommunication, Xi'an 710061, China;

3 School of Electronic & Control Engineering, Chang'an University, Xi'an 710061, China)

Abstract: It is an effective method of building high-performance computing environment by integrating PC clusters on local area network. The definition of cluster abstract system structure and performance model are the kernel problems in realize the aim. In this paper, a performance-analyzing model is put forward for scalable cluster. Using large scale Monte Carlo simulation application as benchmark, the effectiveness of model is proved. The researching results can be used in organizing and planning cluster system.

Key words: scalable cluster; master/worker (m/w) computing framework; SPMD application; grain of parallel task

1 引言

动态可扩展机群系统是无共享的、基于 Internet 或 Intranet 的, 具有开放系统结构及协议标准的机群系统, 突出表现为计算资源动态可用性 & 异构性两大特点. 要使用动态可扩展系统, 系统编程范型需要强调四个指标: 可编程性、适应性、可靠性和有效性^[1].

互联机器的集合并不能成为一个单一计算机系统. 典型的多计算机系统为主从结构, 根结点为主控机(master), 一组子结点为工作结点机(worker). 系统以 master 为代表和访问点, 具体功能的实现由 master 分配任务给各 worker 完成. 这里, 为了方便

叙述, master 和 worker 分别指完成相应功能的机器和进程.

在 m/w 模型中, 当 worker 数目增加时, master 可能成为整个并行系统的性能瓶颈, 限制系统的可扩展性. 为此, 可以从系统结构及并行算法上进行多方面探索, 利用间接方式提高模型的有效性, 而如何根据应用问题确定合理的 worker 数目, 成为有效使用 m/w 模型的关键.

2 动态可扩展机群系统性能属性

按照并行计算 PCMA 算法, 并行计算过程包括任务划分 P(partitioning)、数据传递 C(communication)、任务映射 M(map)、结果聚集 A(agglomera-

收稿日期: 2009-08-16; 修回日期: 2009-10-16

基金项目: 陕西省自然科学基金项目(2009JM8011)

tion) 四步. 在动态可扩展机群环境中, 利用 m/w 编程范型, 将可并行计算部分划分为若干个子任务, 并将完成子任务所需控制体嵌入在线程体中, 通过多线程创建, 实现多 worker 对多个任务的并行执行, 其算法流程如图 1 所示.

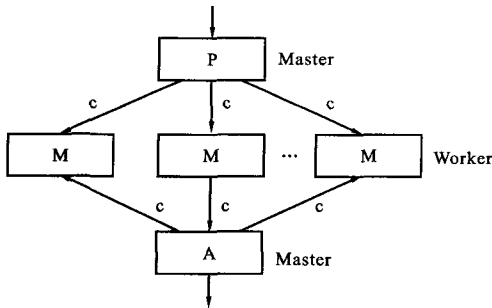


图 1 并行计算 PCMA 流程图

在上述计算模型中, master 是关键部件, 它通过创建和启动多线程完成对远程计算产生延迟的隐藏, 实现系统并行处理功能. 对于一个给定应用, master 可有效控制和管理的并发线程数是有限的, 因此, m/w 计算模型下, 机群系统所具有的并发性不因 worker 多少而决定. 事实上, worker 数目与 worker 利用率及整个应用运行时间之间具有三种可能区域^[2].

(1) 不饱和区: worker 数目少, 每个 worker 利用率高, 整个应用运行时间长, 增加 worker 可以明显改善应用运行时间;

(2) 饱和区: worker 数目适当, 具有最短运行时间和最佳 worker 利用率;

(3) 超饱和区: worker 数目过多, 运行时间有很小的缩短, 而 worker 利用率显著下降.

因此, 对于一个具体应用, 确定合理的 worker 数目是 m/w 计算模型有效运行的关键.

为了简化问题分析复杂性, 假设所用 worker 为同构系统, 且整个计算可以划分为具有相同控制功能和大小的任务, 即属于 SPMD (single program multiple data) 编程问题. 不同子任务具有相同数据结构的人口参数, 完成计算并获得执行结果, 在网络和系统处于轻负载假设下 (即系统除 SPMD 应用外, 没有其它应用在运行), 各个子任务执行时间相等.

定义模型分析所用参数.

t_0 : 数据传输的启动延迟;

λ : 单字节传输延迟, 即带宽的倒数;

v_s : 完成一个子任务所需发送参数的字节数, 由于是 SPMD 问题, 每个任务需要发送数据量相同;

v_r : 子任务完成后返回的计算结果字节数;

t_s : 远程过程调用参数传输延迟, 包括启动延迟 (t_0)、发送延迟 (v_s/λ) 及线路传输延迟 (线路长度 / 传输速度). 由于在 Intranet 环境下, 网络规模有限, 在此忽略延迟, 因此

$$t_s = t_0 + v_s/\lambda;$$

t_r : 远程计算结果返回延迟. 与 t_s 的含义类似, 即

$$t_r = t_0 + v_r/\lambda;$$

t_m : 完成一次并行任务执行需要的通信开支,

$$t_m = t_s + t_r;$$

t_c : 单个子任务串行运行时间;

T_1 : 给定任务串行运行时间;

T_p : 给定任务并行运行时间;

p : m/w 机群系统并行 worker 规模.

机群系统中, worker 之间通过 LAN 组网, 不同 LAN 之间通过各种互联 (集线器, 路由器或网关) 设备, 利用 TCP/IP 协议形成机构内部 Intranet, 同时, 利用路由器与外界 Internet 连接. 目前, 连接单机系统的 LAN 多为 Ethernet. Ethernet 具有总线型拓扑结构, 采用 CSMA/CD 协议解决通信信道访问冲突, 当网络通信负载增大时, 有效带宽大大下降. 同时, 为保证计算参数传递的正确性, 高层采用面向连接的 TCP 协议, 因此, master 与 worker 之间建立连接时, 通信启动延迟 (t_0) 包括逻辑信道建立过程, 主 CPU 执行的通信启动程序需要调用网络控制器 (网卡) 和传输介质来完成通信连接建立过程; 而收/发数据的传输延迟部分 (v_s/λ , v_r/λ) 为利用网络有效带宽由网卡控制完成. 因此, 文献 [1, 3] 在假设完成远程计算各部分可严格并行执行情况下, 即通信启动过程 (t_0) 由 CPU 控制; 数据传输 (v_s/λ , v_r/λ) 由网卡控制; 并行计算 (t_c) 由远程 worker 完成, 建立了机群系统性能分析模型. 由于模型对传输延迟及启动延迟完全并行化利用, 对于一个具体应用, 模型分析的系统规模及加速效果与实际测试结果之间有较大出入.

事实上, 完成一次并行计算的启动过程和数据传输过程都是由 master 主线程通过对网卡的分时使用完成的, 网卡成为两个处理过程中的临界资源, 因此, 利用传输延迟创建新线程体并不能获得真正的并行性. 在此, 我们将发送延迟 (t_s) 和接收延迟 (t_r) 作为一个整体, 不再进行组成部分内部的并行化使用.

定义 1 并行计算任务粒度(β)为完成相应计算任务所需的计算时间与通信开支的比,即

$$\beta = t_c / t_m$$

定义 2 机群 m/w 计算模型并行加速比(S_p)为完成给定计算任务串行计算时间与并行计算时间的比例,即

$$S_p = T_1 / T_p$$

定义 3 资源利用率(E_p)为机群 m/w 计算模型下,给定工作结点数目(p),多 workers 完成计算任务的效率,在此,利用系统实际加速比与理想加速比的百分数描述,即

$$E_p = (S_p / p) * \%$$

设 master 主线程(main)创建多子线程时,按照创建顺序,先创建线程具有较高优先级,忽略线程创建及任务分配所需时间,完成一次远程过程调用的时间包括通信开支和实际计算时间,即($t_m + t_c$),其中,计算部分在远地工作结点机上执行,而通信延迟部分需要经过 master 单一网卡及网络接入线路完成数据传递过程,针对具体并行任务的粒度属性(即,任务不同部分执行时间比例)和任务不同部分执行设备, master 的主控线程在第一个并行计算线程开始远程计算后,利用计算延迟创建和启动其他并发线程.如果并行计算任务粒度为 β ,在远地计算 t_c 延迟内可以创建和启动并行任务数为 β ,按照每个远程计算执行时间相差($t_s + t_r$)原则,可以在不同 workers 与 master 之间形成并行流水线,其具体实现方法为:在每个远程计算线程首次启动时引入一定延迟(t_r),以后多线程计算、传输形成一个最大规模为($\beta + 1$)的并行任务处理流水线,如图 2 所示.

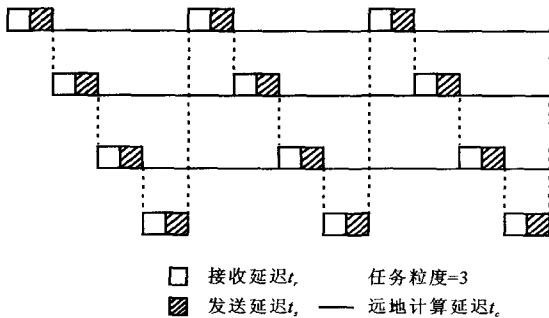


图2 同构机群 m/w 计算模型并行任务流水线

因此,针对该具体应用,机群系统最大规模($P_{\max}$)具有以下属性:

$$P_{\max} = \beta + 1 \quad (1)$$

在相应系统规模下,机群系统加速比为

$$S_{P_{\max}} = p * t_c / (t_m + \beta * t_m + t_c + t_r) \quad (2)$$

当 β 比较大时, $S_{P_{\max}}$ 约为 $\beta/2$, 系统计算资源利用率 $E_{P_{\max}}$ 接近 50%.

同时,当并行任务数(n)大于系统规模时,可以通过多次循环($m, m = n / P_{\max}$)完成整个任务,相应系统具有的加速比和资源利用率为

$$S_{P_{\max}} = \frac{n * t_c}{m * (t_m + t_c) + t_c + t_r} \quad (3)$$

简化后,得

$$S_{P_{\max}} > \frac{m * \beta + m}{2m + \beta + m * \beta} * \beta \quad (4)$$

$$E_{P_{\max}} = \frac{S_{P_{\max}}}{P} = \frac{m + m * \beta}{2m + \beta + m * \beta} > 50\% \quad (5)$$

对于给定 β , 循环次数越多, 加速效果越好.

当系统实际规模小于最大规模时, 相应的加速比和系统计算资源利用率为

$$S_p = \frac{p * t_c}{(p * t_m + t_c + t_r)} \quad (6)$$

简化后,得

$$S_p > \frac{p * \beta}{p + \beta + 1} \quad (7)$$

$$E_p = \frac{S_p}{p} = \frac{\beta}{p + \beta + 1} \quad (8)$$

当并行任务数(n)大于系统规模时, 多次循环运行, 机群系统加速比、效率为

$$S_p = \frac{n * t_c}{m * (t_m + t_c) + (p - 1) * t_m + t_r} \quad (9)$$

$$S_p \approx \frac{m * p * \beta}{m * (\beta + 1) + p} \quad (10)$$

$$E_p = \frac{S_p}{p} = \frac{m * \beta}{p + m * (\beta + 1)} \quad (11)$$

基于以上分析模型, 当为一个具体应用组织机群系统时, 可以通过对并行计算任务粒度的分析, 设计合理的系统规模.

3 性能模型实验分析

利用实验室同构机群(Pentium IV, 2.0GHz)进行模型有效性验证. 结点机之间利用 100Mb/s Ethernet 集线器连接, 通过代理连入校园网. 关闭系统内其它应用, 则网络及每个工作结点机具有轻负载.

以股票期权 100 万次 Monte Carlo 仿真定价算法作为测试程序, 用不同计算粒度机群进行运行效果统计^[4]. 该应用实例的实际并行运行时间、加速效果及机群性能与模型分析结果对比可获得以下研究结论:

(1) 随着并行任务粒度增加, 实际运行效果与模型分析结果更加接近;

(2) 更小的粒度划分将无法获得稳定运行时

间,说明在机群环境下创建高性能计算环境,它所支持的计算必须是粗粒度(即高计算/通信比)的;

(3) 当传输延迟比较大时,由于系统的开放性,并行计算任务完成时间将会有很大的变化,实际性能会与理论分析结果之间存在较大出入。

无论是同构或异构机群系统,都会受动态随机因子的影响,有效利用资源还需要配合有效的任务分配及负载均衡算法^[5-6]。

4 结束语

机群系统的性能不是简单的机器性能的累加,针对具体应用,合理规划机群规模,是提高系统资源利用率及响应速率的关键。本文结合同行任务粒度,定义了 m/w 计算模型机群系统性能分析模型,可以有效实现上述目标。

参考文献:

- [1] Chun Che Fung, Ja Bin Li, Kok Wai Wong, et al. A Java - based parallel platform for the implementation of evolutionary computation for engineering applications[J]. International Journal of Systems Science, 2004, 35(13/14):

741 - 750.

- [2] Heymann E, Senar M A, Luque E, et al. Efficient resource management applied to master - worker applications [J]. Journal of Parallel & Distributed Computing, 2004, 64(6): 767 - 773.
- [3] Anna Morajko, Eduardo Cesar, Paola Caymes - Scutari. Development & turning framework of master/worker applications[J]. JCS&T, 2005, 5(3):115 - 120.
- [4] Luis Paulo Santos, Vitor Castro, Alberto Proenca. Evaluation of the communication performance on a parallel processing system[J]. Lecture Notes In Computer Science, 1997(1332):41 - 48.
- [5] 兰蓉, 郑守淇, 桂小林. 基于局域网机群系统的并行 Monte Carlo 仿真环境的建立[J]. 微电子学与计算机, 2005, 22(3): 96 - 98.
- [6] 亓雪冬, 李霞, 梁鸿, 等. 计算网格中动态负载均衡策略研究[J]. 微电子学与计算机, 2009, 26(5):75 - 77.

作者简介:

吴潜蛟 男, (1964 -), 硕士, 副教授, 中国计算机学会高级会员. 研究方向为计算机网络与通信技术、智能交通控制系统。

(上接第 82 页)

调制器的静态功耗小于 $80\mu\text{W}$. 仿真结果显示在温度信号带宽内, 调制器信噪比为 93.6 dB, 很好地吧低频处的噪声及失调整形到带外. 最后对调制器进行了版图设计, 整个版图所占面积 $620\mu\text{m} \times 312\mu\text{m}$.

参考文献:

- [1] Schreier R, Temes G C. Understanding delta - Sigma data converters[M]. Piscataway, NJ: IEEE Press, 2005.
- [2] 沈佳铭, 洪亮, 石春琦, 等. 一种可重构的 24bit $\Sigma - \Delta$ 调制器的设计[J]. 微电子学与计算机, 2007, 24(4): 159 - 162.
- [3] 徐景, 莫亭亭, 方乐章, 等. 用于 D 类放大器的 $\Delta\Sigma$ 调制器系统分析与设计[J]. 微电子学与计算机, 2008, 25(2): 164 - 167.
- [4] Norsworthy S R, Schreier R, Temes G C. Delta - sigma data converters theory, design and simulation[M]. New York: IEEE Press, 1997.
- [5] Bakker A, Thiele K, Huijsing J H. A CMOS nested chopper instrumentation amplifier with 100-nV offset [J]. IEEE Journal of Solid - State Circuits, 2000, 35(12): 1877 - 1883.
- [6] Enz C C, Temes G C. Circuit techniques for reducing the

effects of op - amp imperfections: autozeroing, correlated double sampling, and chopper stabilization [J]. IEEE, 1996(84): 1584 - 1614.

- [7] Geerts Y, Steyaert M, Sansen W. Design of multi bit delta-sigma A/D converters[M]. Norwell, MA: Kluwer, 2002.
- [8] Lampinen H, Vainio O. An optimization approach to designing OTAs for low - voltage sigma - delta modulators [J]. IEEE Transactions on Instrumentation and Measurement, 2001, 50(6): 1665 - 1671.
- [9] Razavi B. Design of analog CMOS integrated circuits[M]. New York: McGraw - Hill, 2000.

作者简介:

姚舜 男, (1984 -), 硕士研究生. 研究方向为模拟集成电路设计。

顾晓峰 男, (1971 -), 博士, 教授, 博士生导师. 研究方向为半导体器件、集成电路设计。

冯向光 男, (1969 -), 博士, 教授, 高级工程师. 研究方向为集成电路设计。

于宗光 男, (1964 -), 博士, 教授, 博士生导师. 研究方向为集成电路设计。