第 33 卷 第 3 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.3 Computer Engineering February 2007
·网络与通信· 文章编号：1000—3428(2007)03—0114—02 文献标识码：A 中图分类号：TP393

网络流量对网上机群计算的影响

胡 凯，赵宗弟，邓 可
(北京航空航天大学计算机学院，北京 100083)

摘 要：利用网上空闲处理机组成动态非专用机群进行分布式并行计算是网络计算的重要研究方向之一，非专用机群面临的一个最重要的问题就是网络流量动态变化和通信延迟造成网络环境的不确定性和计算的不稳定。该文从理论上分析和评估了网络流量对非专用机群计算环境下各种并行任务的影响，并给出了典型的实验结果。

关键词：分布式并行；机群；网络流量；通信延迟

Influence of Network Flow to the Cluster Computing on LAN

HU Kai, ZHAO Zongdi, DENG Ke
(School of Computer Science & Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100083)

【Abstract】 It is currently an important research direction of network computing to build a dynamic non-dedicated cluster by using idle processors on a LAN. The most important issue of non-dedicated is the dynamic variety of network flow and communication delay. It is a major problem to induce the computing instability. This paper tries to analyze theoretically and estimate the influence of network flow to the various parallel tasks running in a non-dedicated cluster. The typical measurement results are provided.

【Key words】 Distributed and parallel; Cluster; Network flow; Communication delay

机群计算系统是目前国内外研究的热点领域，一般遵循两个研究方向[1]：(1)由一组同构的工作站或微机通过高速网络连接构建的专用机群；(2)非专用机群，它不是由专用的处理机组成，而是通过收集网上空闲处理机组成。据资料统计，许多网上工作站和微机的利用率都小于 10%，因此，自然想到要利用这些闲散的CPU处理能力，这也称之为CPU周期窃取技术，应用程序可以窃取网上处理机中空闲CPU时间。这种方式符合网络共享的特点，方便灵活且经济实惠，是较有前途的一种网络分布式并行计算模式。这种模式更符合当前网络的架构和应用方向，且更为贴近普通用户。但由于当前的网络架构，非专用机群面临的问题和技术困难远高于专用机群，其中一个最重要的问题就是网络流量不稳定和延迟造成网络环境的不确定性，形成非专用机群的瓶颈，本文试图从理论和实验上评估网络流量对非专用机群计算环境下各种并行任务的影响。

1 网络分布式并行计算模式

这里考虑常见的利用网上空闲处理机组成非专用机群进行分布式并行计算的模式，其基本组件包括机群用户主机、机群节点机、通信协议和驻留在各机中的中间件软件，机群用户机也可以是网上机群服务器，主要功能包括：建立动态机群和结点动态信息资料库，接受用户任务，根据用户提交的任务和资源状况决定任务的分配和调度，并监视各结点任务执行情况。机群成员结点机是收集来的网络上匿名空闲处理机，它采取本地任务优先原则。它的主要功能包括：负责接收和监视远程任务的执行，并和机群服务主机进行交互协调，收集本地尽可能多的动态变化信息，周期性地送往主机。组件间协议族定义组件间对话过程的规则说明，包括通信的规约、公共变量和数据结构等。同时，网上还包括大量的其它正工作的各种机器，空闲机本身也并非完全空闲，其本地工作任务也是变化的，其工作模式如图 1[2]所示。因此，网络流量和通信延迟是十分复杂的，也会极大地影响机群中相互协同的分布式并行任务的执行。

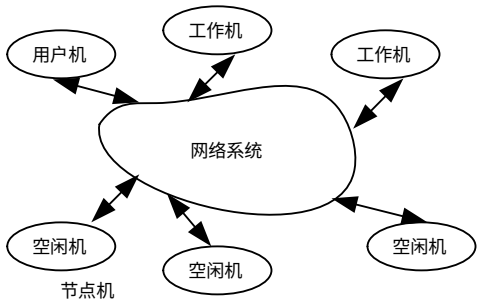

图 1 网上分布式并行计算环境

2 网络流量模型

如果机群构筑在通用网上，在目前条件下，网络的阻塞和延迟是影响系统的最大问题，其影响与具体的网络环境和当时的网络流量有关，其变化范围可能较大，无法得出定量的一般性算法。在特别拥塞的网路上，机群实际上是无法正常发挥效能的。这里将对影响机群计算的通信流量因素进行形式化分析，并在特定系统环境下进行了一些仿真实验，以说明和评估网络流量带来的影响[4,5]。

由于这种机群由网络上分散的用户处理机组成，因此网络流量对系统的影响主要来自两方面。

(1)机群内流量：当机群内处理机之间协同进行并行计算时，产生任务之间的工作消息流量，同时由于其自治性，拥

作者简介：胡 凯(1963－)，男，博士、高工，主研方向：分布式系统和网络计算；赵宗弟，硕士生；邓 可，博士生、高工
收稿日期：2006-02-25 **E-mail：**hukai@buaa.edu.cn

有者可能还在进行其它的一些工作或网上访问，产生与并行任务无关的本地消息流量，导致竞争。

(2)机群外流量：未参与机群的网上其它计算机也会同时进行着网上交互，从而可能导致网络拥塞。

如图 2 为非专用机群的一般通信模型。

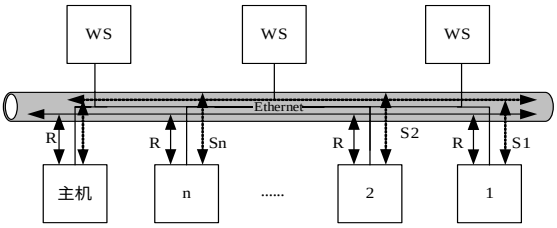

图 2 通信延迟分析模型

假定机群中有 1～n 个空闲的结点处理机，网上还有其它若干正工作的计算机 WS。图 2 中实线代表机群内并行任务间消息通信，虚线代表结点机本地用户消息流量。

定义 1 消息传输率 为某消息流平均每秒传送消息的个数。则网络的理想消息传输率为

φ＝网络带宽/(消息平均长度×8)

定义 2 有效通信比η：

η＝实际消息传递个数/最大消息传递个数

实际网络传输率为η×φ，η的值取决于机群外和机群内并行任务消息流量，机群外流量是实时变化的，只能根据 t 时刻的网络状态测得，这里用系数ξ(t)表示这一部分影响。在并行任务间周期性消息交互时，机群内并行任务间的有效通信比δ(m)是可以求得的，它与消息流个数 m 有关。

η＝ξ(t)δ(m)

定义 3 通信延迟率：

D＝((TD－TN)/TN)*100%

其中 TN 为理想状态下的通信时间，TD 为实际通信时间。假设并行任务间消息流发送率为 R，本地用户消息流发送率为 Si，有 M 个消息要传送。

3 流量影响分析

在非专用机群系统中并行任务间通信可能有 3 种情况[3]。

(1)多对一通信：如n个结点向主机发送消息，同时有本地消息流，共有n个消息流，在每个结点i，消息传输率为R＋Si。在没有本地消息流和拥塞的情况下，每一个传输流的速率是φ/n，则结点i发送M个消息的时间为TN_i＝M×n/φ。拥塞可能发生在n个消息流竞争主机输入链路以及工作消息流和本地消息流在各结点上竞争输出链路。此时平均通信时间为

$$TDi=\frac{M}{\frac{R+Si}{\sum_{j=1}^{n}Si+nR}\xi(t)\delta(n)\varphi(\frac{R}{Si+R})}$$

$$=\frac{M}{\varphi\xi(t)\delta(n)}(\frac{1}{R}\sum_{j=1}^{n}Sj+n)$$

则通信延迟比为

$$D=\frac{\max_{i=1}^{n}TDi-\max_{i=1}^{n}TN_i}{\max_{i=1}^{n}TN_i}$$

$$=\frac{1}{n\xi(t)\delta(n)R}\sum_{j=1}^{n}Sj+(\frac{1}{\xi(t)\delta(n)}-1)$$

上述公式说明对于给定的 R，本地消息流的和越大，通信延迟越大，而本地消息流之和一定时，工作消息流 R 越大，则延迟越小。

(2)一对多通信：如主机发送消息到其它 n 个处理机，接收结点机存在工作消息流和本地消息流的竞争，在没有本地消息流和拥塞的情况下，主机发送消息到各处理机的时间为 TN=nM/R，有本地流和拥塞时通信时间为

$$TDi=\frac{M}{\xi(t)\delta(2)\varphi(\frac{R}{Si+R})}$$

$$=\frac{M}{\varphi\xi(t)\delta(2)}+\frac{MSi}{\varphi\xi(t)\delta(2)R}$$

则通信延迟比为

$$D=\frac{\sum_{i=1}^{n}TDi-TN}{TN}=(R+S)\frac{1}{\varphi\xi(t)\delta(2)}-1$$

其中 S 为 Si 中最大者。这种情况下，并行任务受 R、S 和的影响，与上面情况不同的是，通信延迟同时正比于 S 和 R。

(3)多对多通信：如n个结点机平均每个送消息到m个特定的处理机，此时，发送机有本地消息流和工作消息流的竞争，接收机有本地消息流和工作消息流的竞争，设结点j一次接收了nj个消息，在没有本地消息流和拥塞的情况下，从结点i到结点j的通信时间TN(i,j)=n_j*M/φ，则结点i串行送消息到m个结点的时间为

$$TN_i=\sum_{j=1}^{m}TN(i,j)=\frac{M}{\varphi}\sum_{j=1}^{m}nj$$

当存在本地消息流和拥塞时，从结点 i 到结点 j 的通信时间为

$$TD(i,j)=\frac{M}{\xi(t)\delta^2(2)\varphi(\frac{R}{Si+R})(\frac{R}{Sj+R})}$$

$$=\frac{M}{\varphi\xi(t)\delta^2(2)}(1+\frac{SiSj}{R^2}+\frac{Si}{R}+\frac{Sj}{R})$$

$$TDi=\sum_{j=1}^{m}TD(i,j)$$

如果假设工作消息流消息数是相同的(设n_j=c)，S为Si中最大者，则通信延迟比为

$$D=\frac{\max_{i=1}^{n}TDi-\max_{i=1}^{n}TN_i}{\max_{i=1}^{n}TN_i}$$

$$=(R+S)^2\frac{1}{R^2\xi(t)\delta^2(2)c}-1$$

给定 R，通信延迟取决于 Si 中最大者。

4 模拟测试

为进一步说明实际的通信延迟，在 4 台PC机组成的非专用机群上进行简单的仿真测试，使用PVM消息传递编程环境[6,7]，编制一个简单的点对点消息传递程序，用于模拟本地消息流，如FTP，Telnet等命令流，消息间隔时间可调，以模拟本地消息发送率。另外模拟两个并行程序：PR1 和PR2。PR1 模拟多到一情况，只包含一次计算和消息传递，即在程序结束时向主机传递结果，这是批处理任务模式 (图 3)。这种模式下，本地消息流的影响不大。PR2 模拟多到多情况，这是并行协同任务模式，PR2 重复进行计算、消息传递(图 4)。当本地消息流较小时，影响PR2 性能不大，当本地消息流增大到一定值时，延迟呈线性增长。

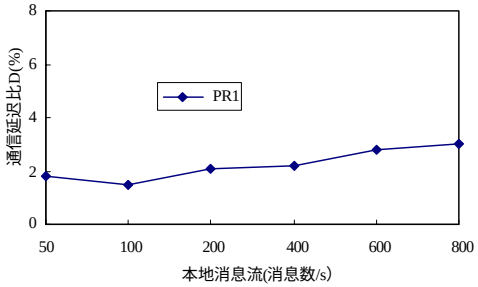

图 3 多对一实验

（下转第 124 页）