LogP简化模型参数估计

刘 辉[1]，戴 波[1]，张 艳[2]，张文彬[3]

(1. 电子科技大学计算机科学与工程学院 成都 610054; 2. 深圳大学信息工程学院 广东深圳 518060;
3. 北京贝尔实验室 北京 100075)

【摘要】针对LogP微观通信模型涉及参数较多，其算法分析较复杂；而简化的LogP模型把两台处理机传送长度为N的消息的所需时间分为：与数据量无关和与数据量相关两部分，从而大大简化了算法分析。通过模拟简化的LogP模型的参数,并在LogP环境下对单机和多机分别进行测试，利用测试数据分析网络延迟和软件开销，得出经验公式，从而证明该模型可以正确而有效简化并行算法的设计和分析。

关 键 词 简化的LogP模型; 参数模拟; 经验公式; 网络通信

中图分类号 TP 331 文献标识码 A

Simple LogP Model’s Parameters Simulate

LIU Hui[1]，DAI Bo[1]，ZHANG Yan[2]，ZHANG We-bin[3]

(1. School of Computer Science and Engineering, UEST of China Chengdu 610054;
2. School of Information and Engineering, SZU of China Guangdong Shenzhen 518060; 3. Bell Lab. Beijing 100075)

Abstract The LogP model has more parameters and more complicated arithmetic. The simple LogP model divided the deferent data to two parts to make algorithmic analyses simple. For simulating simple LogP model’s parameters, we supply one method to test idiographic environment’ s parameters of model, which can help us to design and analyze parallel arithmetic. And we respectively use the single machine and multiply machines to test the model ’s parameters of LogP. According to the test data, we analyzed the network delay and the software expense. Then we get an empirical formula for simple LogP model’s parameters.

Key words simple LogP model; simulate parameters; experiential formula; network communication

1 基于LogP模型的通信模型分析

LogP模型是当前最具实际意义的并行计算模型。LogP模型有4个参数(L，o，g，P)，其中，L，g反映了互连网络传送消息的平均固有性能，o反映了处理机对网络协议的处理能力，P为处理机数。

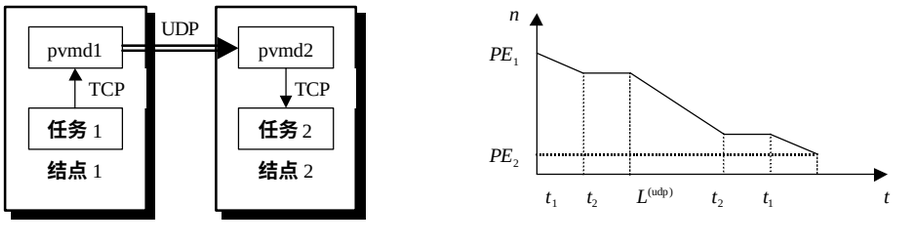
下面将LogP模型应用到通过以太网相连的实际工作站环境中。图1(a)所示为并行虚拟机(Parallel Virtural Machine，PVM)系统中位于两个不同处理机之间典型的消息传递方式。首先，任务(Task)1通过面向连接的传输控制协议(Transfer Control Protocol，TCP)将消息传给本地监控进程pvmd1，并继续自己的工作，不再管理消息在网络中的传输；此后，pvmd1通过面向用户数据报协议(User Data Packet, UDP)，将从task1中接收到的消息传送给远程监控进程pvmd2；最后，当远程任务task2向pvmd2访问该消息时，进程pvmd2再利用面向连接的TCP协议将消息传送给任务(Task)2。至此，一次完整的消息传递过程结束。

收稿日期：2003 – 09 – 15
作者简介：刘 辉(1970 –)，男，讲师，主要从事网络计算、数据库方面的研究.

图1(a)和1(b)分别是消息传递方式和传递一个短消息在各个阶段所花费时间的示意图。假设$L^{(\mathrm{udp})}$与$o^{(\mathrm{udp})}$分别为网络延迟和网机接口系统开销时间，$L^{(\mathrm{tcp})}$与$o^{(\mathrm{tcp})}$分别为本地TCP连接中传递一个短消息所需的延迟时间和处理机花费的CPU时间。轻载条件下，两台处理机传送一个长度为n的消息所需的时间为：

$$t_{n,2}=2t_{n,2}^{(\mathrm{tcp})}+L^{(\mathrm{udp})}+(n+1)\max(o^{(\mathrm{udp})},g)$$

式中 $t_{n,2}^{(\mathrm{tcp})}=L^{(\mathrm{tcp})}+no^{(\mathrm{tcp})}$ 为本地TCP连接传送所花费的时间。



(a) 消息传递方式　　(b) 消息传递花费时间

图1　基于LogP模型的PVM通信模型

对于不同长度的消息，通过研究表明，$L^{(\mathrm{udp})}$与$L^{(\mathrm{tcp})}$随消息长度的增大而略有升高，但当消息量不大(≤16 K)时，增加的幅度不太显著，因此可以将$L^{(\mathrm{udp})}$与$L^{(\mathrm{tcp})}$这两个参数固定为传送几百字节时的延迟时间。同时，将o，g的单位统一为μs/B(称为单字节开销)，则$o^{(\mathrm{udp})}$与$o^{(\mathrm{tcp})}$的单字节开销随消息长度的增大而减小，反映为网络频带的提高。而LogP模型中的L和o反映的是互联网络的平均性能。

推广到含有 P 台工作站的网络环境，假设某次消息交换要求第 i $(1\le i\le P)$台处理机分别发送和接收长度为 NS_i 和 NR_i 的消息，则要求流经网络的总消息量为 $N=\sum_{i=1}^{P}(NS_i+NR_i)$。由于以太网总线机制的共享介质特性，同一时刻至多允许有一个数据报文流经网络。因此，轻载条件下，可认为所有信息是按串行方式流经整个网络，所需时间为：$t_{N,P}^{(\mathrm{comm})}=t_{N,P}^{(\mathrm{tcp})}+L^{(\mathrm{udp})}+No^{(\mathrm{udp})}$。其中 $o^{(\mathrm{udp})}$ 按每台处理机发送和接收的消息长度来选取(或取平均值)$t_{N,P}^{(\mathrm{tcp})}=2L^{(\mathrm{tcp})}+(\max_i\{NS_i\}+\max_i\{NR_i\})\times o^{(\mathrm{tcp})}$。而第 i 台处理机花费的CPU时间为：

$$t_{N,P_i}^{(\mathrm{cpu})}=2L^{(\mathrm{tcp})}+(NS_i+NR_i)(o^{(\mathrm{tcp})}+o^{(\mathrm{udp})})$$

考虑一般情况，$NS_i=NR_i=N/P(1\le i\le P)$，$N$ 充分大时，得：

$$\frac{t_{N,P}^{(\mathrm{comm})}}{t_{N,P_i}^{(\mathrm{cpu})}}=\frac{2L^{(\mathrm{tcp})}+L^{(\mathrm{udp})}+2N/Po^{(\mathrm{tcp})}+No)^{(\mathrm{udp})}}{2L^{(\mathrm{tcp})}+2N/P(o^{(\mathrm{tcp})}+o^{(\mathrm{udp})})}\approx\frac{2o^{(\mathrm{tcp})}+Po^{(\mathrm{udp})}}{2o^{(\mathrm{tcp})}+2o^{(\mathrm{udp})}}\approx\frac{P}{2}$$

即 $t_{N,P}^{(\mathrm{comm})}\approx\frac{P}{2}t_{N,P_i}^{(\mathrm{cpu})}$，可见在通信量较大时，网络总线机制的效率很低。而在采用交换技术的网络中，通信效率就会提高很多。

2　简化的LogP模型

由于LogP微观通信模型涉及参数较多，基于LogP微观通信模型的算法分析比较复杂，所以推出了把两台处理机传送长度为 N 的消息所需的时间分为两部分的模型，即与数据量无关部分和与数据量相关部分的简化的LogP模型。这样可大大简化分析，即 $T=T_{SR}+T_WN$。式中，T_{SR} 为处理机发送和接收时间之和，它与所传送的数据量大小无关，也称为发送启动时间；而 T_W 是单位字节传递时间，T_WN 是与数据量相关的那部分时间；N 是发送的字节数。这种推广模型也适用于长消息包的发送/接收。

下面就针对简化的LogP模型进行参数模拟，验证简化的LogP模型的有效性和正确性，从而帮助算法进行设计与分析。

3　LogP模型参数模拟

以太网是共享介质的总线传输机制，所有处理机在网络中的位置等价，且轻载时任意两台处理机间传递相同长度消息的时间是相同的，所以可用点对点的通信来确定L和o。

对点对点通信，采用“乒乓法”进行测试，即进程0发送长度为 N 字节的消息到进程1，然后等待从进程1返回的消息；此时，进程1执行一个阻塞式接收语句，一旦收到进程0发送来的消息，立即返回一个同样的消息。将该过程重复进行多次，排除起始两次通信，取平均值再除以2，就得到点对点通信所需的时间。变换消息长度，得到通信时间，然后采用一阶线性拟合的方法，就得到点对点通信所需时间的近似经验公式。测试环境是由四台配置完全相同的微机(均为K6/266，32M RAM)组成的10 M局域网，软件平台为SCO UNIX OpenServer System 5.0.2和PVM 3.3.11。记一个数据包的软件开销和网络传输延迟分别为 $\hat{o}$ 和 $\hat{L}$，而该数据包平均每字节的软件开销和网络延迟分别为 o 和 L。

3.1　单机模拟

由于单机没有网络传输延迟(单机情况下的延迟主要来自于$L^{(\mathrm{udp})}$)，可以把单机情况下从进程0发送数据到进程1接收到数据的通信时间表示为 $T_{2\hat{o}}$，测试结果如表1所示。

计算得出线性相关系数 $r_{2\hat{o}}$=0.999 7，所以通信时间 $T_{2\hat{o}}$ 与数据量 N 呈显著线性相关，拟合出单机情况下点对点通信的经验公式 $T_{2\hat{o}}=(3.384+0.756N)$ μs。在这个公式中，认为常数3.384是与所传送的数据量大小无关的部分(启动时间)，而带有系数0.756的部分是与传送数据量大小有关的部分。

3.2　多机模拟

在多机网络情况下的延迟主要来自于$L^{(\mathrm{udp})}$和$L^{(\mathrm{tcp})}$。可以把多机情况下从进程0发送数据到进程1接收数据的通信时间表示为 $T_{2\hat{o}+\hat{L}}$，测试结果如表2所示。做一阶线性拟合。计算得出线性相关系数 $r_{2\hat{o}}$=0.989 1，所以通信时间 $T_{2\hat{o}+\hat{L}}$ 与数据量 N 也呈显著线性相关，所以拟合出在多机网络情况下点对点通信的经验公式：$T_{2\hat{o}+\hat{L}}=(123.345\ 9+2.101N)$ μs。同样，这个公式可分成与数据量无关部分和与数据量相关部分两部分。

表1　点对点通信单机模拟结果

数据量/K B	通信时间/s	网络延迟/μ s · B^{-1}
64	0.055	0.839
128	0.100	0.763
256	0.190	0.725
512	0.380	0.725
1 024	0.750	0.715
2 048	1.570	0.749

表2　点对点通信多机模拟结果

数据量/K B	通信时间/s	字节开销/μ s · B^{-1}
64	0.155	2.365
128	0.295	2.251
256	0.590	2.251
512	1.155	2.203
1 024	2.290	2.184
2 048	4.450	2.122

3.3　网络延迟与软件开销的分析

将 $T_{2\hat{o}}$ 与 $T_{2\hat{o}+\hat{L}}$ 的拟合公式相减，得 $T_{\hat{L}}$ 的拟合公式 $T_{\hat{L}}=(119.961\ 9+1.345N)$ μs。

在单机和多机情况下，可以分别求出单字节的传输时间为：

$$T_{2o}=\frac{T_{2\hat{o}}}{N}=\frac{3.384}{N}+0.756\xrightarrow{N\to\infty}0.756\ \mu\mathrm{s}$$

$$T_{2o+L}=\frac{T_{2\hat{o}+\hat{L}}}{N}=\frac{123.346}{N}+2.101\xrightarrow{N\to\infty}2.101\ \mu\mathrm{s}$$

所以，可以求出网络延迟 L 和软件开销 o，即：

$$o=\frac{T_{2o}}{2}=\frac{1.692}{N}+0.378\xrightarrow{N\to\infty}0.378\ \mu\mathrm{s}$$

$$L=\frac{T_{\hat{L}}}{N}=\frac{119.962}{N}+1.345\xrightarrow{N\to\infty}1.345\ \mu\mathrm{s}$$

可以看到，当 N 充分大时，字节传输时间趋近于一个常数。在表1、表2中可以看出字节开销稳定地逼近于一个常数。通过表1、表2还可以求出单字节的网络延迟和软件开销，如表3所示。通过表3可以看出网络延迟和软件开销也趋近于一个常数(并不和传输的数据量 N 成线性关系，其相关系数分别为 r_L=0.125，$r_o=0.379$)，而且其值与上面通过模拟公式求出的值相符合。所以，在某个范围内，可以认为单字节网络延迟和软件开销为一个常数，例如在实验局域网环境中，网络延迟为1.345 μs，而软件开销为0.378 μs。

表3 网络延迟和软件开销的模拟结果

数据量/KB	网络延迟/μs·B^{-1}	软件开销/μs·B^{-1}
64	1.526	0.419
128	1.488	0.382
256	1.526	0.363
512	1.479	0.363
1 024	1.469	0.358
2 048	1.373	0.375

另外，随着传输数据量的增加，每字节所需的软件开销便逐渐减小，所以在松散网络结构中大粒度是获取高效率的必要手段。

4 结束语

通过提供的LogP简化模型具体环境中的参数模拟，针对单机和多机的模拟结果，分析了其网络延迟和软件开销，得到其经验公式。因此可以确定LogP参数在该环境中的简化是正确而有效的。因此简化的LogP模型可以有效而正确地简化实际并行算法的设计与分析。

参 考 文 献

[1] 莫则尧, 李晓梅. 工作站网络环境下的并行计算[J]. 计算机学报, 1997, 20(6): 34 -37

[2]Matthew I, Agarwal A. LoPC: modeling contention in parallel algorithms[C]. ACM 0-89791-906-8/97/0006. Portland, Oregon, United States, 1997. 56-69

[3] Keeton K, Patterson D A, Anderson T E. LogP Quantified: the case for low-overhead local area networks [C]. In Hot Interconnects III, San Francisco, California, United States, 1995. 82-84

[4] Alexandrov A. Ionescu M F. LogGP: incorporating long messages into the LogP model for parallel computation [C]. Journal of Parallel and Distributed Computing 44, Ottawa, Canada, 1997. 36-39

[5] 孙家昶, 张林波. 网络并行计算与分布式编程环境[M]. 北京: 科学出版社, 1996

编 辑 熊思亮

(上接第212页)

参 考 文 献

[1] Kunz T. The influence of different workload descriptions on a heuristic load balancing scheme[J]. IEEE Transactions on Software Engineering, 1991, 17(7): 725 -730

[2] Pai V S, Aron M, Banga G, et al. Locality-aware request distribution in cluster-based network servers[C]. In Proceedings of the Eighth International Conference on Architectural Support for Programming Languages and Operating Systems(ASPLOS-VIII), San Jose, California, 1998

[3] Aron M, Druschel P, Waenepoel W Z. A mechanism for resource management in cluster-based network servers[C]. In Proceedings of the ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems, Santa Clara, USA, 2000

[4] Castro M, Dwyer M, Rumsewicz. M Load balancing and control for distributed world wide web servers[C]. In Proceedings of the 1999 International Conference on Control Applications, Kohala Coast, Hawaii, 1999

[5] Liu Xin-song. The performance research of the distributed parallel server system with distributed parallel I/O interface[J]. ACTA Electronica Sinica, 2002, 30(12): 1 808-1 811

编 辑 漆 蓉