基于Myrinet的用户空间精简协议*

董春雷 郑纬民

(清华大学计算机科学与技术系 北京 100084)

E-mail: zwm-dcs@tsinghua.edu.cn

摘要 通信子系统是影响工作站机群系统整体性能的主要因素.文章在分析和比较了3种常用的网络性能之后,指出上层协议的处理是影响工作站机群系统性能的主要瓶颈.在由640Mbps的Myrinet连接的8台Sun SPARC工作站组成的机群系统上实现了一个用户层的高性能的精简通信协议——RCP(reduced communication protocol).通过精简协议的冗余功能、减少数据拷贝次数和直接操作硬件缓冲区等方法,达到低延迟、高效率.RCP的回路延迟时间比TCP/IP小得多(200μs vs 1 540μs),可用带宽也高得多(178Mbps vs 34Mbps),协议的带宽利用率达到80.5%,并为上层应用程序和PVM提供了一个简单、容易的接口.

关键词 工作站机群,精简通信协议,并行处理,PVM.

中图法分类号 TP393

通信系统是并行工作站机群系统的重要组成部分.它实现系统的消息传递功能,其性能的好坏是影响并行计算加速比和效率的主要因素之一[1,2].实现一种快速的消息通信机制是提高工作站机群系统性能的有效方法[3,4].

(1) 通信系统的现状

① 高速网络发展迅速

由于包括并行计算在内的多种应用对高速网络的需求,推动了网络技术的飞速发展.目前出现了多种新型的高速网络,如Fast Ethernet,ATM,Myrinet,这些新型网络的速度是传统Ethernet的10倍或更高.由于高速网络的运用,使得影响通信系统性能的瓶颈已从过去的网络硬件转移到网络通信软件上,也就是说,对高速网络系统而言,网络链路的传输延迟已经很小,而严重影响通信系统性能提高的却是通信协议的处理开销.通信协议的处理开销过大,阻碍了高速网性能的大幅度提高.Myrinet的物理链路带宽为640Mbps,而在TCP/IP协议层测得的带宽只有42Mbps.可见,上层通信协议的开销使高速网络的高性能得不到充分的发挥.

② 峰值带宽增长幅度大,但往返延迟改进小

所谓高速网络性能的提高,往往是指其峰值带宽.与传统Ethernet相比,高速网的峰值带宽有了很大程度的提高,从10Mbps增长到100Mbps,640Mbps,甚至1Gbps.网络的带宽反映的是大数据包的传输能力.衡量并行系统的通信性能的另一个重要参数是往返延迟,它决定了系统的计算粒度,直接影响系统的并行效率.这些高速网络的峰值带宽增加了,应用程序的可见带宽也会有较大的提高,但在实际系统的测试中发现,往返延迟的改善不大.表1是我们对10Mbps Ethernet和640Mbps Myrinet一个字节数据包往返延迟时间的测试结果(在Sparc 20工作站上).

从表1中看出,Myrinet的通信延迟反而比普通Ethernet还大.可见,高速网络在传输小数据包时的性能并不理想,而小数据包下网络系统性能的好坏,对并行计算有很大影响,因为用户设计的并行算法经常需要进行小

---

* 本文研究得到国家863高科技项目基金资助.作者董春雷,1964年生,博士,主要研究领域为并行处理,机群系统,新型网络协议.郑纬民,1946年生,教授,博士生导师,主要研究领域为并行处理.

本文通讯联系人:董春雷,北京 100084,清华大学计算机科学与技术系

本文1997-10-09收到原稿,1998-03-17收到修改稿

表 1 Ethernet 与 Myrinet 网络延迟时间

网络类型	TCP 往返延迟(μs)	UDP 往返延迟(μs)
Ethernet	1 438	1 146
Myrinet	1 506	1 370

数据量的数据交换,而且并行计算环境,如 PVM,在运行时也需发送很多数据量很小的控制消息进行系统运行情况的监护.对其他方面的应用而言,小数据包的传输延迟也起到举足轻重的作用.例如,UC Berkeley 通过对 NFS 数据包的跟踪发现,95%的数据包的大小不超过 192 字节,数据包的平均大小是 382 字节[5].因此,减小高速网络小数据包传输的延迟时间是提高网络系统性能的重要方法之一.

③ 通信协议基本不变

新型网络技术的涌现,并没有对传统的通信协议进行较大的改进.即使像 ATM 这种完全不同于 Ethernet 的高速网络,也是想方设法地使用 TCP/IP[6]协议,其中一个重要的原因在于产品的兼容及推广,因为 TCP/IP 协议几乎占据了网络通信的各个领域.但对并行机群系统来说,兼容性不是它的主要考虑因素,因此,可以对传统的通信协议机制、软件实现作较大的修改,克服传统协议的弊病,使高速网络的优越性得以充分展现.

(2) 影响通信系统性能的因素

前面的分析说明,限制高速网络性能的瓶颈是软件上的问题,下面从软件的角度分析影响通信系统性能的主要因素.软件引起的通信开销主要由协议处理和操作系统处理两方面的开销造成.操作系统提供了网络协议的实现环境,也决定了网络协议的某些实现方法,网络协议本身的处理方式也都是影响通信延迟的因素.下面对这两种因素进行较为详细的分析.

① 传统的 TCP/IP 协议各层次重复实现的功能很多

TCP/IP 协议是面向低速率、高差错和大数据包传输而设计的,是一个多层次的软件结构.协议层次多,使得对数据的操作和拷贝次数增多,引起延迟时间的增加.另外,在多层协议的实现中,各层还重复实现了很多相同的功能,比如:(a) 从链路层到传输层都要进行差错控制;(b) 从链路层到应用层都要进行连接建立和释放;(c) 从网络层到应用层都要进行协议的处理机调度;(d) 从链路层到应用层都要进行流量控制;(e) 从网络层到应用层都要进行组装和定序的缓冲.

这些冗余的功能虽然可以确保数据的无差错传送,但随着链路传输出错率从 10^{-4}降至 10^{-9},这种冗余处理反而限制了数据及时提交给应用程序处理.可见,多层次的协议结构是造成通信瓶颈的主要原因之一,合并某些层次,删除冗余的处理,设计一种轻型通信协议,是提高通信性能的重要方法.

② 协议复杂的缓冲管理增加了网络延迟

网络协议处理包括很多功能,如流量控制、差错控制、出错重发机制、拥塞控制等,而这些功能的实现都与缓冲管理密切相关.

缓冲管理的作用是完成数据的分组和组装,缓冲区可看成一种网络资源,这种资源是有限的,对它的管理很重要.不过,通常的缓冲管理机制都比较复杂.例如,Berkeley UNIX 采用一种叫做 mbufs 的结构对协议的数据包进行缓冲管理,但 mbufs 的算法很复杂,开销很大.在 DECstation 5000 上,对单字节 TCP 消息缓冲管理,mbufs 需要 100μs,而对 512 字节数据包需要 300μs[5],可见,缓冲管理带来的网络延迟也很大.如何简化协议复杂的缓冲管理也是我们研究的主要内容.

③ 操作系统额外开销不可忽视

操作系统提供的系统调用和原语是网络协议实现的底层软件支持.在网络协议实现中涉及到上下文切换、调入/调出页面、启动 I/O 设备、中断响应等操作系统处理,有时这些开销可能比协议本身的处理开销还大.比如,在 Sun 3/60 系统上,TCP/IP 对一个数据包的协议处理时间为 100μs,而操作系统的额外开销却高达 240μs,这就造成了通信性能对操作系统一定程度上的依赖.目前,一种广泛采用的思想是不修改 UNIX 操作系统,而在用户空间实现一个用户态的协议层,并使此协议层能够旁路操作系统的影响,直接对网络硬件设备进行操作.这样就可减小这部分的开销,最大限度地提高通信系统的性能.

为了减小通信协议处理开销,我们实现了一个基于 Myrinet 的通信协议 RCP(reduced communication protocol),其基本思想是抛弃传统的 TCP/IP 协议,尽量以底层通信接口为基础,简化协议的层次和不必要的环节,达到减少拷贝次数、简化缓冲管理和降低协议处理开销的目的.

1　精简通信协议 RCP 的结构

精简通信协议 RCP 是一个建立在 Myrinet 的 API 层上的用户空间协议,它为应用程序提供有序、可靠的双向传输.一端用户空间的数据通过一条虚拟通道直接拷贝到远程的用户空间中,不需要系统的多次缓冲与干预,RCP 的结构如图 1 所示.

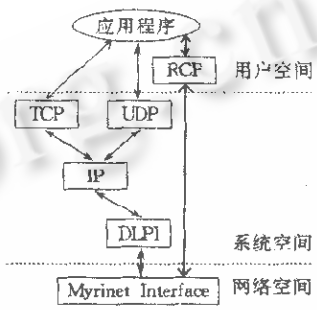

图1　一般协议与精简通信协议RCP的结构

RCP 的设计目标是要减小网络的回路延迟,尽量提高链路层的带宽利用率.它建立在 Myrinet 的 API 层上,以充分利用 Myrinet 的硬件性能,为应用程序提供对网络硬件的直接控制.

RCP 设计的另一个目标是要实现一个简单、容易的用户编程接口.目前,工作站机群系统上的并行程序开发环境是 PVM,RCP 直接通过 PVM 的函数接口来实现,这样,应用程序只要作很少的改动就可在精简协议上执行 PVM 并行程序.RCP 协议和 PVM 结合起来,不仅简化了用户接口,而且减少了 PVM 的一次数据拷贝,节省了 PVM 的内部缓冲内存的要求,为大数据量应用程序提供更大的内存,从而降低了访盘次数,提高了效率.

2　RCP 的实现

RCP 的基本思想是在两个用户进程之间提供一个顺序、可靠的虚拟通道.它是一个平衡协议,发送方和接收方完全对等,无论是首先执行发送例程,还是首先执行接收例程,都会在某一点同步并交换数据,它类似于 PVM 中的阻塞式发送接收.

2.1　RCP 中包的流程

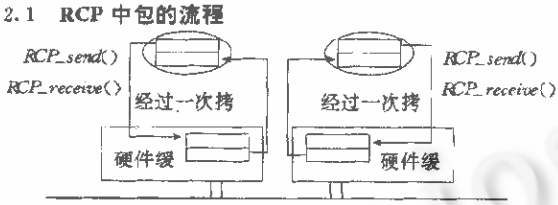

图2　RCP的通信流程

RCP 的通信流程如图 2 所示.

Myrinet 层允许的最大包长是 8 432Bytes,但是,RCP 的传输数据长度只受可分配内存空间的限制.数据的打包与拆包直接在用户提供的原缓冲空间中进行,这样既降低了系统缓冲需求,又减少了数据拷贝次数.

2.2　顺序性、可靠性

为了取得较高的效率,同时基于测试的结果,RCP 中只对包长和包头信息进行检查,而不对数据部分计算校验和,因为 Myrinet 接口卡具有检错能力,并能保证首先发送的包先到达目的地.数据的顺序性由一个基于连接的全局计数来维护,它对于发送和接收是完全对称的.其可靠性是由一个简单的滑动窗口协议和选择重传机制来保证.窗口大小可以根据硬件缓冲区的大小和网络负荷的轻重而改变.RCP 的滑动窗口协议,如图 3(a)所示,由发送双方约定,收发双方首先同步,确定窗口大小,然后发送方发送一个窗口的数据,接收方进行确认后,再进行下一个窗口的传送.如果有出错包,则在应答中指明出错包的个数及序号,下一个窗口立即重发出错包,这样,窗口的尺寸越来越小,总能完成传输,如图 3(b)所示.

2.3　连接建立及管理

RCP 是一个对等的协议,它的连接管理采用静态方式,每一通道的两端进程都要维护这一连接,因而这种联接并不真正要与对方交换连接信息,只是在本地作一记录,为以后的发送及接收提供必要的信息.

2.4　开始与结束

在实际的应用程序中,可能不知道进程之间的准确时序关系,即不知道是首先执行到接收原语,还是首先执

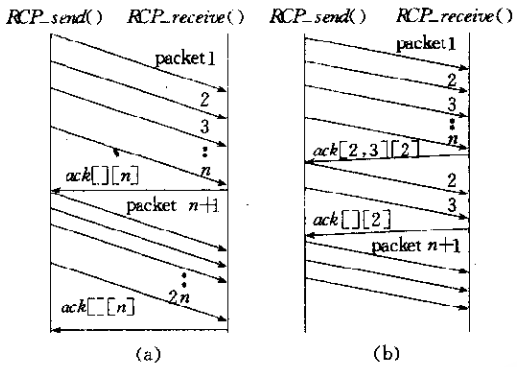

图3 简单的滑动窗口协议

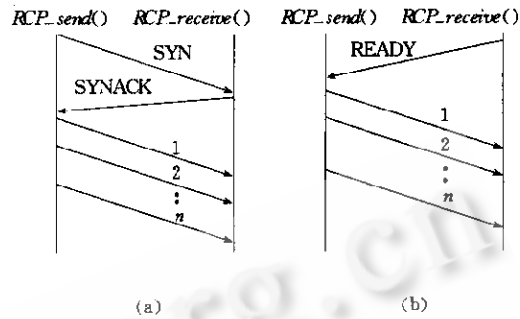

图4 RCP的发送与同步

行到发送原语.对这两种情况,RCP 都可以工作.当先执行到接收原语时,接收方向发送方发一同步信息:READY 包,等待接收数据;发送方根据该信息确定窗口大小并发送数据.如果先执行到发送原语,则分为两种情况:一种是发送大量数据(数据长度>1 个包长),发送方向接收方发一同步信息,然后进入发送阶段;另一种是发送小量数据(数据长度<1 个包长),同步信息和数据一起发出,接收方一起确认.这样可以减小小数据包的传输延迟.不管是哪种情况,如果没在某指定时间内收到应答,则发送方认为发送失败,重新发送,如图 4 所示.

3 RCP 与并行开发环境的接口

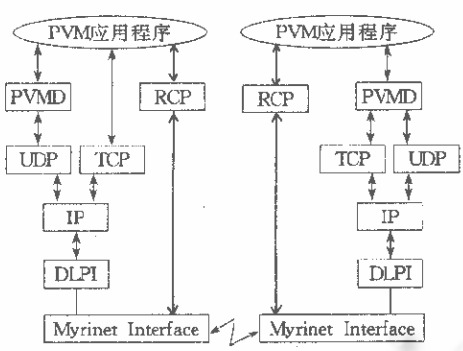

图5 RCP与PVM及应用程序的关系

为了尽量减少应用程序的修改,我们对并行程序开发环境 PVM[7]的部分通信例程进行了重写,直接在原 PVM 的 *pvm_psend*()和 *pvm_precv*()中增加一条新的数据通道(如图 5 所示),从而简化了应用程序的接口.PVM 应用程序不需要显式调用 RCP 的发送接收函数,只是在每个通信进程的开始调用一下 RCP 协议初始化例程,并打开通信通道.RCP 中维护一张从 PVM 任务标识到通信通道的转换表及通道的连接信息.

另外,RCP 还为了和 *pvm_precv*()对应,提供了一个不经过 PVM 缓冲的多播函数 *pvm_pmcast*(),它直接利用 RCP 将用户数据广播到指定任务.

4 RCP 的性能分析

卡耐基-梅隆大学的 Jose C. Brustoloni[8]在 10Mbps 的 Ethernet 上改进了 TCP/IP 协议的实现,使往返延迟减小到 750μs.Utah 大学的 Mark Swanson[9]提出了一种基于发送方的协议,改善了往返延迟和同步.RCP 在清华大学计算机科学与技术系的并行工作站机群系统(8 台 Sun Sparcstation 20/50+640Mbps Myrinet)上进行了测试.RCP 的回路延迟为 200μs,是原 TCP/IP 协议的 1/7;应用程序可见带宽达 178Mbps,是 TCP/IP 带宽的 4 倍,如图 6 所示.

5 结 论

本文对工作站机群系统的通信网络性能进行了测试和分析.基于其结果,在 Myrinet 的 API 层设计并实现了一个建立在用户空间的精简通信协议 RCP,减少了数据拷贝次数和缓冲需求,充分利用硬件功能,具有低回路延迟、高带宽利用率和方便的编程接口等优点.使用 RCP 通信的 PVM 应用程序可以明显提高性能.

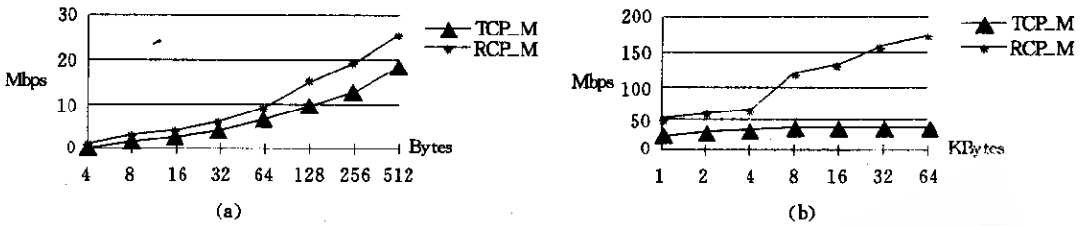

图6　Myrinet上TCP和RCP的带宽对比

参考文献

1 David Culler *et al*. LogP: toward a realistic model of parallel computation. In: Proceedings of the Principles and Practice of Parallel Processing. http://now.cs.berkeleg.edu/paper/logp.ps, 1993. 1～12

2 Thekkath A, Nguyen T D, Moy E *et al*. Implementing network protocols at user level. In: Proceedings of SIGCOMM '94. Sept. 1994. 14～23

3 Felderman, DeSchon A, Cohen D *et al*. ATOMIC: a high-speed local communication architecture. Journal of High Speed Networks, 1994,3(1):1～30

4 Dong Chun-lei, Zheng Wei-min, Wang Ding-xing *et al*. A scalable parallel workstation cluster system. In: Proceedings of APDC'97. Los Alamitos: IEEE Computer Society Press, 1997. 307～313

5 Thekkath C, Levy H. Limits to low-latency communications on high-speed networks. ACM Transactions on Computer Systems, May 1993,11(2):179～203

6 Postel B. Transmission Control Protocol. RFC 792, Sept. 1981

7 Beguelin, Dongarra J, Geist A *et al*. PVM: experiences, current status and future direction. In: Proceedings of the Supercomputing'93. Los Alamitos: IEEE Computer Society Press, 1993. 765～766

8 Jose C Brustoloni, Brian N Bershad. Simple protocol processing for high-bandwidth low-latency networking. Technical Report, CMU-CS-93-132

9 Mark Swanson, Leigh Stoller. Low latency workstation cluster communications using sender-based protocols. Technical Report, UTAH-CS, 1994

A Reduced User-level Communication Protocol for Myrinet

DONG Chun-lei　ZHENG Wei-min

(*Department of Computer Science and Technology　Tsinghua University　Beijing*　100084)

Abstract　Communication subsystem is a major factor that affects the overall performance of a workstation cluster system. In this paper, the authors analyzed and compared the performance of three popular networks, and pointed out that the high level protocol processing is the bottleneck of communication in workstation cluster system. A user level high performance protocol, called RCP(reduced communication protocol), had been built on a cluster of 8 Sun SPARC workstations connected with 640Mbps Myrinet. Low overhead and high efficiency of the protocol are achieved by simplifying the protocol's redundant functions, reducing the times of data copying, and operating directly on hardware buffer. The round trip latency of RCP is much lower than TCP/IP (200μs vs 1 540μs), the application available bandwidth is much improved(178Mbps vs 34 Mbps) and the bandwidth usage of Myrinet on RCP reaches to 80.5%. A very simple interface to application or PVM is provided as well.

Key words　Workstation cluster, reduced communication protocol, parallel processing, PVM.