用户级通信协议 BCL-3 对 IP 协议支持的研究*

陈志辉[1+], 马 捷[2], 陈国良[1], 高 帆[2]

[1](中国科学技术大学 计算机科学与技术系,安徽 合肥 230027)

[2](中国科学院 计算技术研究所 国家智能计算机研究开发中心,北京 100080)

Study on IP Supporting over User-Level Protocol BCL-3

CHEN Zhi-Hui[1+], MA Jie[2], CHEN Guo-Liang[1], GAO Fan[2]

[1](Department of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

[2](National Research Center for Intelligent Computing Systems, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: 86-551-3603145; 86-10-62617948, Fax: 86-551-3601013, E-mail: zhchen@mail.ustc.edu.cn

http://www.ustc.edu.cn

Received 2002-05-11; Accepted 2002-09-17

Chen ZH, Ma J, Chen GL, Gao F. Study on IP supporting over user-level protocol BCL-3. *Journal of Software*, 2003,14(9):1629~1634.

http://www.jos.org.cn/1000-9825/14/1629.htm

Abstract: In order to efficiently exploit high performance network, researchers have developed many user-level protocols, which can achieve high bandwidth and low latency the bottom hardware supplies. But user-level protocols supply a completely different API, which makes them only support science computing, and traditional Socket-based network application program with kernel-level protocol cannot run on them. To solve this problem, a IP supporting module has been implemented on user-level protocol BCL-3, which makes it support both science computing and existing TCP/IP-based network application programs efficiently. And based on software overhead analysis of TCP/IP, BCL-3 adopts some optimization in IP supporting module, which makes TCP/IP over BCL-3 achieve high performance. The improved BCL-3 has been run on Dawning3000L super server. On Dawning3000L, TCP/IP over BCL-3 has achieved performance with maximum bandwidth of 938Mbps and minimum one-way latency of 48.1μs.

Key words: high performance network; user-level protocol; kernel-level protocol; BCL-3; TCP/IP

摘 要: 为了充分利用高性能网络,研究人员开发了多种用户级通信协议.这些用户级通信协议可以获得底层硬件提供的高带宽、低延迟.然而由于它们提供完全不同的应用程序接口,用户级通信协议往往只能支持科学计算,而不能支持传统的基于 Socket 接口、采用核心级通信协议的网络应用程序.通过增加一个 IP 协议支持模块,BCL-3 用户级通信协议在支持科学计算的同时,可以有效地支持现有的基于 TCP/IP 协议的网络应用程序.而

* Supported by the National High-Tech Research and Development Plan of China under Grant Nos.863-306-ZD01-01, 2001AA 111041 (国家高技术研究发展计划(863))

第一作者简介: 陈志辉(1976－),男,江西泰和人,博士,主要研究领域为高性能网络,分布式系统.

且在分析 TCP/IP 协议软件开销的基础上,IP 协议支持模块有针对性地采用了一些优化技术,使运行在 BCL-3 上的 TCP/IP 协议可以取得很高的网络性能.改进的 BCL-3 已经运行在曙光 3000L 超级服务器上.在曙光 3000L 上,运行于 BCL-3 之上的 TCP/IP 协议取得了最大带宽 938Mbps,最小单向延迟 48.1μs 的性能.

关键词: 高性能网络;用户级通信协议;核心级通信协议;BCL-3;TCP/IP

中图法分类号: TP 文献标识码: A

近年来,机群系统逐渐成为高性能计算的主流平台.它们被广泛应用到科学工程计算、事务处理、互连网信息服务上.在机群系统中,通信性能往往成为决定机群系统整体性能的一个关键因素.越来越多的机群系统采用了高性能网络,如 Myrinet[1],这些网络具有高带宽(>1Gbps)、低延迟(<1μs)的高性能特征.然而核心级通信协议,如 TCP/IP 协议,每次通信操作都涉及到多次进出操作系统核心,这使协议的软件开销成为通信的主要开销.为此,研究人员开发了多种用户级通信协议,如 AM[2],BIP[3],BCL-3[4]等.这些用户级通信协议实现了用户进程直接访问网络设备,从而可以充分利用网络硬件提供的高性能.

用户级通信协议一般提供自己专有的 API(应用程序接口),这完全不同于核心级通信协议采用的 API.程序员一般通过与用户级通信协议绑定的上层通信软件(MPI 和 PVM)来使用它,这就限制了用户级通信协议只能应用到科学计算上.

然而近来的研究表明,机群系统只有 10%用来作科学计算,其余 80%~90%都用于事务处理和信息服务.现有的绝大部分网络应用程序,比如 WWW,NFS,FTP 等,都是基于 TCP/IP 协议的.为了在不重写这些应用的基础上,让它们可以得到近似用户级通信协议的高性能,我们在 BCL-3 用户级通信协议上实现了一个 IP 协议支持模块,它能以二进制兼容的方式支持传统的基于 TCP/IP 协议的网络应用程序,同时对 TCP/IP 协议进行了优化,并获得了很高的性能.

本文介绍了在 BCL-3 用户级通信协议上进行的支持 IP 协议的研究.第 1 节分析 TCP/IP 协议的软件开销.第 2 节详细介绍 BCL-3 对 IP 协议支持的研究以及针对 TCP/IP 协议采用的一些优化技术.第 3 节在曙光 3000L 平台上测试了 BCL-3 上 TCP/IP 协议的性能.最后对全文进行总结,并提出了下一步的研究计划.

1 TCP/IP 协议的软件开销分析

TCP/IP 是一个多层次的核心级通信协议,每个层次分别负责不同的通信功能[5].这导致 TCP/IP 协议在应用于高性能网络时,软件的开销成为其性能瓶颈.下面我们将具体分析 TCP/IP 协议可能存在的开销.在第 2.3 节中,针对这些软件开销,我们还提出并实现了在高性能网络上减少这些开销的优化技术.

1.1 内存拷贝

TCP/IP 协议是一种核心级通信协议,这样,在一次消息传递过程中,需要涉及消息在发送方从用户空间到核心空间以及在接收方从核心空间到用户空间的数据拷贝.这对于大消息的传递来说,会是一个很大的开销.

1.2 数据校验

TCP/IP 协议在传递一个消息时,消息的发送方和接收方都需要对消息进行数据校验和的计算,以保证消息传递的正确性.在某些 TCP/IP 协议的实现中,当消息足够大时,这种计算所耗费的 CPU 时间可以达到整个消息发送和接收所耗用 CPU 时间的 46%[6].

1.3 硬件中断

TCP/IP 协议也是一个异步通信协议,一般采用中断机制来实现.网络设备驱动程序通过中断方式来通知一次数据发送和接收的完成.在操作系统中,中断是一项比较耗时的操作,因为处理一次中断需要涉及进程上下文的切换.

1.4 数据包的分片和重组

在 TCP/IP 协议中,消息的大小受 MTU(最大传输单元)的限制.当前以太网的 MTU 为 1 500 个字节,这导致在发送大消息时,IP 协议在发送方需要把大消息分成一个个小的数据分片进行发送,而在接收方,IP 协议必须对接收到的数据分片进行重组,以形成一个完整的消息.IP 协议的分片和重组对大消息的传递是一个比较大的开销.

2 BCL-3 对 IP 协议支持的研究

2.1 BCL-3通信协议

BCL-3 是国家智能计算机研究中心为曙光 3000 系列超级服务器开发的一种用户级通信协议,它可以充分获得底层硬件提供的高性能.BCL-3 通信协议由 3 部分组成:用户级基本通信库、运行于操作系统核心的设备驱动程序以及运行于 Myrinet 网卡上的 MCP(Myrinet control program).用户级基本通信库包括了 BCL-3 提供给用户的 API,设备驱动程序负责管理全局数据结构,向 MCP 寄送操作请求,实现虚地址到物理地址的转换等,而 MCP 程序解释并执行来自主机方的命令,控制数据在主机与接口之间以及接口与网络之间的传送.

BCL-3 采用基于端口的消息传递机制.在 BCL-3 中,进行消息通信的基本单位是端口,每个进程可以创建一个端口与其他进程进行通信,并由该进程所在的节点号与其创建的端口号所标识.每个端口包括一个消息发送请求队列、一个消息接收缓冲池和相应的完成事件队列:消息发送完成事件队列和消息接收完成事件队列.消息接收缓冲池由一组消息通道组成,BCL-3 定义了 3 种通道类型:系统消息通道、一般消息通道、开放消息通道.其中系统消息通道用于小消息的接收,另外两种消息通道用于接收大消息.

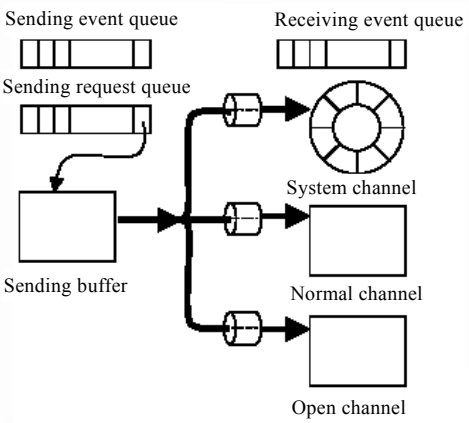

Fig.1 BCL-3 communication mechanism
图 1 BCL-3 通信机制

图 1 给出了 BCL-3 通信机制的示意图.当一个进程发送一个消息给另外一个进程时,消息发送进程首先需要获取并填写一个消息发送请求.消息发送请求中包含了待发送消息的数据或其缓冲区地址、消息接收进程的标识(节点号和端口号偶对)以及目标通道.然后,消息发送进程将该消息发送请求插入消息发送请求队列中.在实施正式的消息传递之前,消息接收进程需要准备好其相应的接收通道.当消息接收完成之后,接收方的 MCP 会产生一个消息接收完成事件,并插入消息接收完成事件队列,以通知消息接收进程.同样,在消息发送完成以后,发送方的 MCP 也会产生一个消息发送完成事件通知相应的进程.

2.2 IP协议支持模块

图 2 表示了增加 IP 协议支持模块后的 BCL-3 通信协议层次.其中,IP 协议支持模块是对 BCL-3 的设备驱动程序模块的扩展,也处在操作系统核心层,可以完全使用 BCL-3 设备驱动程序管理的全局数据结构.

IP 协议支持模块保留 BCL-3 的最后一个端口作为 IP 专用端口.IP 端口不同于普通的 BCL-3 端口,它在 IP 协议模块加载时被分配并初始化,不专属于某个进程,而且采用中断机制来完成消息的接收.在 IP 端口中,系统消息通道被用来接收小消息,而一般消息通道被用来接收大消息.所有的一般消息通道被虚拟成一个 IP 通道,内部组织为一个循环队列,并预先分配操作系统中网络子系统的专有缓冲区作为其接收缓冲区.

IP 协议支持模块在操作系统核心注册一个以太网络设备驱动程序模块,从而可以实现和上层 IP 协议的交互.IP 协议支持模块接收到来自上层 IP 协议的数据包以后,获取并填写一个发送请求,然后把该发送请求插入到消息发送请求队列中去,由 MCP 去负责完成发送.发送请求包括目标节点号、端口号以及目标通道等,其中目标节点号通过 IP 协议到 BCL 协议的转换机制获取,端口号为 IP 端口,目标通道根据消息的大小确定是系统消息

通道还是 IP 通道.在接收方,MCP 收到一个发往 IP 端口的消息后,把该消息 DMA(direct memory access)到目标通道的接收缓冲区中去,产生一个消息接收完成事件,并插入消息接收完成事件队列,最后发一个中断信号给主机.IP 协议支持模块处理该中断,通过查看消息接收完成事件队列,把目标通道接收缓冲区中的消息提交到上层 IP 协议的接收队列中去,如果该消息是通过 IP 通道接收的,则重新为 IP 通道中对应的一般用户通道分配一个网络子系统的专有缓冲区.

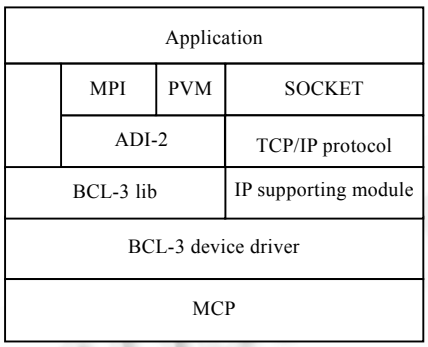

Fig.2 Layers of BCL-3 protocol
图 2 BCL-3 通信协议层次

2.3 IP协议支持模块采用的优化技术

在第 1 节对 TCP/IP 协议软件开销分析的基础上,IP 协议支持模块有针对性地采用了一些优化技术,从而尽可能地降低 TCP/IP 协议的软件开销,提高其性能.这种优化完全在 IP 协议模块中实现,并不涉及对 TCP/IP 协议或者 Socket 接口的修改.IP 协议支持模块中针对 TCP/IP 协议采用的优化技术主要有:

(1) 减少数据校验和计算的次数.Myrinet 网卡实际上已经在链路层上具有完备的数据校验功能,BCL-3 的 MCP 接收到数据校验出错的消息,会自动丢弃该消息.因此,IP 协议支持模块认为每个接收到的消息都是正确的,并在消息的头部设置一个标志,使 TCP 协议忽略对该消息进行数据校验和的计算.

(2) 减少中断次数.在 IP 协议支持模块中,我们尽可能减少中断的次数.发送方完成一次消息发送后,并不向主机发送中断信号,而是在发送完成事件队列中插入一个发送完成事件.IP 协议支持模块通过访问发送完成事件队列,就可以知道相关的发送请求已经完成,从而释放相应的系统资源.在接收方,IP 协议支持模块采用了中断压缩技术,在一次中断处理过程中,如果还有后继的消息进来,中断处理函数就继续处理后继的消息.因而在一次中断处理过程中,IP 协议支持模块可以接收多个消息.

(3) 更大的 MTU.Myrinet 交换机采用直通路由机制,这使 Myrinet 网络上传输的消息大小不受交换机缓冲区大小的限制,从而 IP 协议支持模块可以选择比较大的 MTU.目前的实现采用 8 192 个字节作为其 MTU,这使得通过 IP 协议支持模块发送大消息的时候,可以极大地减少在 IP 协议层的数据分片和重组.

(4) 带宽和延迟的综合考虑.BCL-3 支持两种访问 Myrinet 卡上内存的方式:DMA 和 PIO(programming IO).它们具有不同的特点,DMA 需要 CPU 的干预比较少,传输数据效率高,但 DMA 的启动开销比较大.而 PIO 是由 CPU 直接发出 move 指令进行数据传输,启动时间比较小,但每次传输都需要占用 CPU 周期.因此,PIO 适合于小消息的传输,而 DMA 适合于大消息的传输.实验表明,在传输小于 128 个字节的消息时,PIO 的性能比 DMA 要好[7].由于 BCL-3 本身受到设计的限制,IP 协议支持模块对小于等于 64 个字节的消息采用 PIO 的方式,而对大于 64 字节的消息,则采用 DMA 方式.这种灵活的实现,使 BCL-3 在传输小消息的时候,可以取得低延迟,而在传输大消息的时候,可以取得高带宽.

3 性能测试

本文在曙光 3000L 超级服务器上测试了 BCL-3 以及百兆以太网的 TCP/IP 协议性能.曙光 3000L 是一个 SMP 机群,每个节点有两个 PIII866CPU 以及 1G 的内存,安装 TurboLinux Server 操作系统,该操作系统使用版本

为 2.2.18 的 Linux 内核.节点之间有两套互连网络:Myrinet 网和百兆以太网.Myrinet 网络采用 M3S-PCI64B-2 网络接口卡及 M3S-SW8 交换设备.为了进行比较,在另外一个采用千兆以太网、其他配置和曙光 3000L 一样的 SMP 机群系统上,我们测试了千兆以太网的 TCP/IP 协议性能.

3.1 单向延迟

延迟测试采用我们自己写的一个测试程序,这个测试程序在两个进程之间以乒乓方式发送固定大小的消息.程序中设置了 Socket 接口的 TCP_NODELAY 选项以禁止 Nagle 算法,从而强制消息尽快被发送出去,这样减少了小消息发送的带宽,但可以取得更好的延迟性能.Socket 的发送和接收缓冲区都被设置为 512K 字节.图 3 是我们测试得到的结果.从图 3 可以看出,BCL-3 的延迟性能明显地要比百兆以太网和千兆以太网的好,而且在消息大小从 4 字节跳变到 8 字节时,BCL-3 上 TCP/IP 协议的单向延迟有一个明显的变化,那是因为在我们的实现中,小于等于 64 个字节的消息采用 PIO 的方式发送,而大于 64 字节的消息采用 DMA 的方式.在这 64 字节中,除去各个协议的头部信息,真正有效的最大消息数据只有 6 字节.

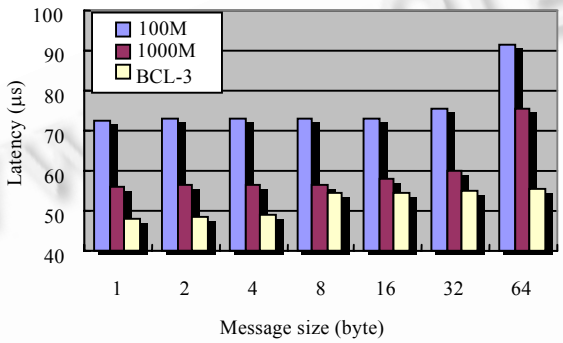

Fig.3 One-Way latency performance of TCP/IP protocol
图 3 TCP/IP 协议的单向延迟性能

3.2 网络带宽

带宽测试采用当前国际上通用的网络性能测试程序 netperf.我们使用了 netperf 的 2.1pl3 版本.在测试过程中,Socket 接口的发送和接收缓冲区设置为 512K 字节,测试模式为 TCP_SREAM.为了了解数据校验对性能的影响,本文测试了 IP 协议支持模块在接收方也进行数据校验的性能.图 4 是我们的测试结果,从图 4 可以看出,当发送的消息足够大时,BCL-3 上 TCP/IP 协议的带宽远远高于千兆以太网的带宽.当发送的消息大小为 8 000 字节时,BCL-3 取得了 938.6Mb/s 的最高带宽.同时也可以看出,有接收方数据校验的 BCL-3 只取得了 748.1Mb/s 的最大带宽.

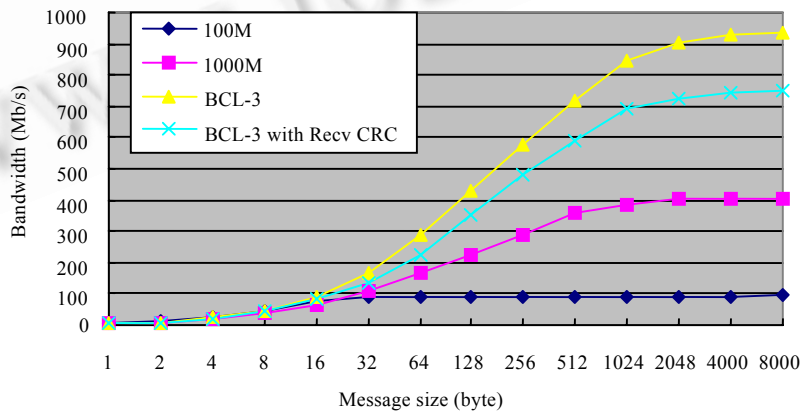

Fig.4 Bandwidth performance of TCP/IP protocol
图 4 TCP/IP 协议的带宽性能

4 结论和进一步研究

通过在用户级通信协议 BCL-3 的设备驱动程序上增加一个 IP 协议支持模块,可以使 BCL-3 在有效支持科学计算的同时,能够以二进制兼容的方式支持基于 TCP/IP 协议的网络应用程序,从而极大地拓宽了 BCL-3 的使用范围.而且,在分析 TCP/IP 协议的软件开销基础上,针对机群系统以及高性能网络的特点,在 IP 协议支持模块中采用一些有效的优化技术,可以在用户级通信协议上取得很高的 TCP/IP 协议性能.

为了让 BCL-3 更加有效地支持 IP 协议,我们将进行后续的研究与开发,研究的重点主要有:在 MCP 上实现 Scatter/Gather 功能以支持 TCP/IP 协议的零拷贝功能;开发 Myrinet 网卡的硬件校验功能,从而减少发送方进行数据校验的开销;考虑在一次消息传递的过程中,MCP 要访问发送完成事件队列和接收完成事件队列这两个数据结构,而 MCP 访问主机方内存只有通过 DMA,这会带来一些额外的开销,可以把这两个数据结构存放在 Myrinet 卡的共享内存上,让 IP 协议支持模块通过 PIO 去访问它们,从而提高 BCL-3 上 TCP/IP 协议的延迟性能.

致谢 特别感谢国家智能计算机研究开发中心孙凝晖、孟丹、肖利民老师对我们研究和工作上的指导,同时感谢他们为我们提供了良好的学习和研究条件.另外,还要感谢丁建峰、霍志刚、刘淘英等同学的热情帮助和支持.

References:

[1] Boden NJ, Cohen D, Felderman RE, Kulawik AE, Seitz CL, Seizovic JN, Su WK. A gigabit-per-second local area network. IEEE Micro, 1995,15(1):29~36.

[2] von Eicken T, Culler DE, Goldstein SC, Schauser KE. Active messages: A mechanism for integrated communication and computation. In: Abramson D, Gaudiot JL, eds. Proceedings of the 19th Annual International Symposium on Computer Architecture. Gold Coast: ACM Press, 1992. 256~266.

[3] Prylli L, Tourancheau B. BIP: A new protocol designed for high-performance networking on myrinet. In: Rolim JDP, ed. Lecture Notes in Computer Science 1388, Orlando: Springer-Verlag, 1998. 472~485.

[4] Ma J. Research on key Issues of communication system on cluster of SMP's [Ph.D. Thesis]. Beijing: Institute of Computing Technology, The Chinese Academy of Sciences, 2001 (in Chinese with English abstract).

[5] Stevens WR. TCP/IP Illustrated Volume 1: The Protocol. Addison-Wesley Publishing Company, 1994. 1~5.

[6] Kay J, Pasquale J. The importance of non-data touching processing overheads in TCP/IP. ACM Sigcomm. Computer Communication Review, 1993,23(4):259~268.

[7] Shen J, Zheng WM, Ju DP. FMP: A fast messages passing for workstation clusters. Chinese Journal of Computers, 1998,21(7):595~602 (in Chinese with English abstract).

附中文参考文献:

[4] 马捷.基于 SMP 节点的机群通信系统关键技术的研究[博士学位论文].北京:中国科学院计算技术研究所,2001.

[7] 申俊,郑纬民,鞠大鹏.FMP:一种适用于机群系统的快速消息传递机制.计算机学报,1998,21(7):595~602.