基于 Myrinet/GM 的多通道通信*

张继超+, 舒继武, 郑纬民, 常 迪

(清华大学 计算机科学与技术系,北京 100084)

Multi-Networking Communication Based on Myrinet/GM

ZHANG Ji-Chao, SHU Ji-Wu, ZHENG Wei-Min, CHANG Di

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

+ Corresponding author: Phn: 86-10-62785592, Fax 86-10-62780969, E-mail: zjc99@mails.tsinghua.edu.cn

http://www.cs.tsinghua.edu.cn

Received 2002-04-03; Accepted 2002-06-12

Zhang JC, Shu JW, Zheng WM, Chang D. Multi-Networking communication based on Myrinet/GM. *Journal of Software*, 2003,14(2):278~284.

Abstract: Communication subsystem is crucial for cluster computing, which affects its efficiency, adaptability and scalability. Large-Scale applications require challenging communication performance and availability from cluster systems. Multi-Networking communication is a novel approach to improve the communication performance and availability by using multiple network links in parallel. In this paper, the effect of multi-process multiplexing one network link is analyzed, a dynamic link dispatch scheme is proposed, and the design and implementation of a multi-networking communication layer, MNC is introduced, which extends GM messaging layer, and supports multi-Myrinet parallel communication. MNC provides multi-process effectively exploiting the raw performance of multi-Myrinet, and improves the communication performance of application layer significantly. Compared with one-way Myrinet/GM environment, the communication bandwidth between MNC processes has increased by 34% on the PC cluster interconnected with 2-way Myrinet.

Key words: multi-networking communication; high-performance communication; communication protocol; MNC; Myrinet

摘 要: 通信子系统对并行系统的计算效率有重要影响,大规模应用对并行平台的通信性能和可用性提出了挑战性的要求.多通道通信技术通过并行采用多路网络链路互连来提高并行系统通信性能和可用性.首先分析了多进程复用网络对通信性能的影响,然后以 Myrinet/GM 网络平台为基础,提出了基于网络接口层的通信链路动态选择与分配策略,设计和实现了支持多路 Myrinet 网络并行通信的协议层 MNC.MNC 支持通信进程平等,充分地利用多路 Myrinet 网络链路资源.在使用 2 路 Myrinet 互连的 PC 机群平台上,MNC 进程间通信带宽相对于单链路提高了约 34%,有效地提高了应用层通信性能.

关键词: 多通道通信;高性能通信;通信协议;MNC;Myrinet

* Supported by the National Natural Science Foundation of China under Grant No.60103019 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.G1998020300 (国家重点基础研究发展规划(973))

第一作者简介: 张继超(1975－),男,湖北汉川人,硕士生,主要研究领域为并行/分布式处理技术,机群通信系统.

中图法分类号: TP393 文献标识码: A

基于商用计算元件的并行分布式系统已经广泛应用于科学计算和商业应用领域.通信子系统对并行系统的计算效率和适用范围都有重要影响.随着多种高性能互连网络如 Myrinet[1],QsNet[2]以及高效用户态通信协议如基于 Myrinet 的 AM[3],BIP[4],GM[5]等的出现,通信系统性能得到很大的提高,已经达到 Gb 级带宽、微秒级延迟的水平.但相对于计算元件,如处理器和存储器技术的迅速发展,通信系统的发展依然滞后.而大规模的通信密集型应用,如实时模拟、能源勘探、天气预报、核物理研究等对并行平台的通信性能要求提高到 10G 字节的带宽水平,远远超过现有互连网络技术所能提供的通信性能.进一步提高并行系统的通信性能对于充分利用最新计算资源、解决挑战性的技术问题都有重要意义.另一方面,新型高性能互连网络硬件价格昂贵,而传统的低效网络系统如 Ethernet 因不能满足实际高性能的通信需要而得不到进一步的应用.有必要采用并行处理的思路,将已有通信资源利用起来,提供现有网络所不能提供的通信性能.

多通道通信技术(multi-networking communication,简称 MNC)是一种并行利用多路网络链路通信,以提高通信性能,突破单链路互连环境下的性能瓶颈的新型技术.多通道通信技术通过增加计算结点之间的网络链路数,直接成倍地增加了应用层可用的物理通信性能,同时,冗余网络链路的增加也为提高通信的可靠性提供了可能.将多网络接口和多网络链路集成化以实现多通道通信,在提高性能的同时降低成本,将成为今后通信技术的发展方向之一.

相对于已经很成熟的单链路互连网络和用户态通信技术而言,针对多通道技术的通信协议、容错机制、性能评测等研究还很少得到关注.目前国际上只在少数超大规模巨型机上研究多通道通信技术,在国内还未见相关研究的报道.本文以 Myrinet/GM 网络平台为基础,设计和实现了一套底层多通道通信层 MNC.MNC 通过扩展适用于 Myrinet 的高效用户态通信协议 GM,提供了对多路 Myrinet 网络互连的支持,并行利用多路 Myrinet 网络提高了应用层通信性能.

本文的实验结果都基于双 Myrinet 2000 网络链路互连的 PC 机群平台,每条网络链路带宽为 2.0Gbit/s,延迟约为 3μs.每台 PC 使用四路 Xeon 700MHz 处理器,带有 1G 内存,运行 Redhat Linux 7.2 操作系统(内核版本为 2.4.7-10smp),在 64bit/66MHz 的 PCI 总线上插有两块内嵌 LANai 9.0 200MHz 处理器的 Myrinet 网卡.

1 多进程复用网络对通信性能的影响

在多用户、多进程通信环境下,网络链路(netlink)和通信接口(NIC)往往由多个通信进程复用,即有多路数据同时在一道网络链路上传输.来自所有通信进程的服务请求构成消息队列,由网络接口顺序处理,导致每一路数据分享硬件网络的通信能力.本文在 PC 机群平台上测试了多进程复用网络对通信性能的影响.图 1 给出了两对通信进程复用一条网络链路通信的示意图,图 2 给出了每一对进程之间的通信性能和单进程单链路通信的性能比较.

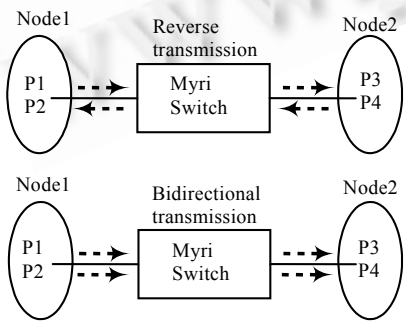


Fig.1 Two transmissions multiplexing one link

图 1 两路数据通信复用单一链路

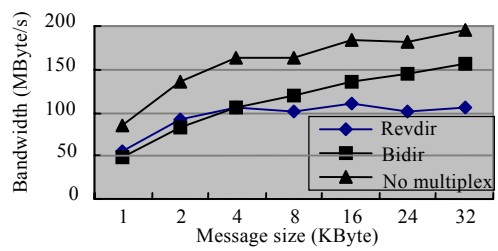


Fig.2 Comparison of bandwidth

图 2 通信带宽比较

从以上测试结果可以看出,在单道网络环境下,当有两路数据同向通信时,每一路数据所利用的有效带宽约为单道网络环境下一路数据通信时带宽的 50%,两路数据正好分别占用一半的网络通信带宽.当有两路数据反向通信时,每一路数据占用的有效通信带宽大约为单路数据通信带宽的 80%,也没有达到 Myrinet 所称的支持全双工链路通信的理论性能.多通信进程对单一网络的共享和复用,导致了每一路数据通信所能利用的通信带宽大大降低.

分析 Myinet/GM 的通信机制可知,通信性能的瓶颈在于网络接口和网络链路本身的物理特点.导致性能降低的原因在于 I/O 总线的共享、网络接口上消息发送和接收 DMA 引擎的相互影响、LANai 接口内部总线的共享和网络链路上数据流的干扰等.这些影响因素是由 Myrinet 网络硬件的特点决定的,无法通过改进协议软件的设计来消除.而采用多通道通信技术既可以增加通信性能,也可以将传统低效互连网络并行起来提供更高的通信性能,从而在降低成本的情况下提供所需通信性能.

2 MNC 的实现与技术

MNC 的设计目标是控制和协调多道 Myrinet 网络向应用层提供高性能通信服务,并向用户提供一个类似单链路网络协议的透明的通信接口.MNC 除了要提供可靠、有序的消息传输和流量控制等通信服务以外,还要尽量提高网络链路的利用率、降低链路的分配和调度开销.MNC 扩展了 GM Library 和 MCP 部分,以控制和协调多路 Myrinet 链路并行通信,并向用户层提供透明的编程接口.

2.1 Myrinet/GM简介

2.1.1 GM 软件结构

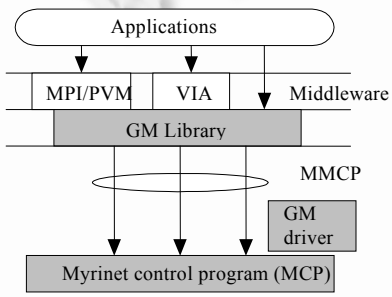


Fig.3 GM architecture
图 3 GM 层次结构

Myrinet/GM 是 Myricom 公司提供的高性能、高可靠、可编程的 Gb 级系统域网络(SAN)及其用户态通信协议,是当前机群系统广泛采用的通信环境.

GM(Glenn's message)软件结构包括 GM 函数库(GM library)、驱动程序(GM driver)和 Myrinet 控制程序(GM MCP)3 个部分,如图 3 所示.GM Library 向用户程序提供编程接口,允许用户程序直接访问网络硬件.GM Driver 运行在操作系统核心,主要负责完成将 MCP 加载到 Myrinet 网卡和初始化,并将 Myrinet 网卡内存空间映射到用户空间.GM MCP 由网络接口处理器 LANai 运行,负责和用户程序以及网络交互,完成数据传输.

2.1.2 GM MCP 结构

GM MCP 的软件结构如图 4 所示.它包括 4 个状态机(也称为引擎):SDMA,RDMA,SEND 和 RECV,它们分别独立地并发执行.各状态机主要作用如下:

SDMA 状态机负责根据用户提交的发送请求,将消息数据从用户缓冲区传送到 SRAM 区域.用户提交的发送请求以发送描述子(send descriptor)的形式传递给 MCP,当 SDMA 发现新的消息描述子后,将该发送请求信息保存到消息发送令牌队列(send token queue)中.此后当 SDMA 检测到该发送令牌时,将消息从用户缓冲区分块转移到网卡传送块队列(transmit chunk)中.

SEND 状态机负责检测网卡传送块队列,在发现有新的传送块时,增加路由信息和有关控制信息后发送到网络上.消息发送完成后,将描述该消息的消息发送

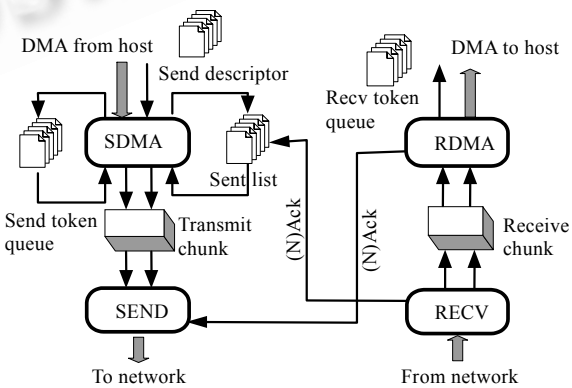


Fig.4 GM MCP architecture
图 4 GM MCP 结构

令牌返回给用户程序继续使用.

在接收端由 RECV 和 RDMA 状态机进行的接收操作与发送操作类似.

2.2 MNC通信缓冲区的管理

在多链路通信环境下,通信进程需要动态选择发送和接收链路.用户 DMA 缓冲区的数据通过任意可用的通信接口传输,从而所有的通信接口都要能够实现 DMA 区域的虚实地址转换.实现该转换有两种途径:一种是由系统调用完成,每次通信前将对应的 DMA 缓冲区物理地址传给当前所用的网络接口;另一种途径是将 DMA 缓冲区的虚实地址转换信息保存到所有网络接口上.前一种转换途径在每次通信前增加了系统调用开销,降低了通信效率.

MNC 是采用后一种方式实现用户空间 DMA 缓冲区的管理.MNC 进程在初始化时同时打开所有可用网络接口上的通信端口,将用户 DMA 缓冲区的虚实地址转换信息保存到所有网络接口上,这样保证了通信进程可以使用本结点的任何网络链路进行通信.MNC 进程通信之前首先要将消息数据拷贝到该 DMA 缓冲区,然后通过动态链路分配策略选择合适的发送和接收链路进行通信,DMA 缓冲区的物理地址只要查询 DMA 区域地址转换表就可以得到.

2.3 MNC网络链路的分配和调度

2.3.1 网络链路分配与调度的一般策略

对多网络链路进行分配和调度是 MNC 设计中的关键部分,它既要保证通信结点的每条链路都能得到充分利用,又要保证每个通信进程平等地分享网络链路资源.网络链路的分配和调度主要有如下 3 种方式[6]:

静态分配策略:通信进程初始化时被静态指定特定的通信链路用于通信.通信时不需要进行链路选择,避免了链路选择开销.但当进程数较多时,会有多个通信进程复用网络,从而导致进程间通信性能的下降.另外,静态分配没有考虑通信进程间实际的通信量,可能某些进程通信量很小,从而导致相应网络链路利用率很低.

单向分配策略:指定结点上某些网络链路用于发送操作,而其他网络链路用于接收操作,每条网络链路的数据流向事先指定.该策略避免了多路数据在一条网络链路上相向传输,但需要进行链路选择,从而引入了一定的额外开销.在通信数据流向比较固定时,会导致某些链路被多进程复用,而某些链路闲置.

动态分配策略:所有的网络链路都可供通信进程发送或接收消息,进程通信之前首先要选择发送和接收链路,尽量避免进程复用网络.这种分配策略考虑到每个进程和链路的通信状态,让每个通信进程尽可能利用空闲的链路通信,避免进程复用网络,同时充分利用链路资源,但引入了一定的链路分配与调度开销.

2.3.2 基于动态分配策略的 MNC 的链路分配和调度

MNC 采用动态分配策略分配进程通信所用的发送和接收链路,对网络链路的选择根据当前的使用状态确定.当 Myrinet 网络接口处于通信状态时,网络接口(NIC)与主机(host)或网络(network)进行数据交换,相应的 DMA 引擎启动 DMA 操作,并设定相应的状态寄存器.GM/MCP 记录的网卡数据传送状态包括 SDMAING(从 Host 到 NIC)、RDMAING(从 NIC 到 Host)、SENDING(从 NIC 到 Network)和 RECVING(从 Network 到 NIC).由于从 Host 到 NIC 的 DMA 操作和从 NIC 到 Network 的 DMA 操作可以采用流水线的方式并发进行,所以只要判断当前网卡是否处于 SDMAING 或 RECVING 状态,就可以判断网卡是否正在进行消息发送或接收操作.

MNC 将网络链路的使用状态区分为 BUSY 和 FREE 两种.当网卡通信状态是 SDMAING 或 RECVING 时,是 BUSY 状态,其他状态则是 FREE 状态.通信进程选择链路时,首先选择处于 FREE 状态的链路,如果当前没有 FREE 状态的链路,则循环选择处于 BUSY 状态的网络链路.

(1) 源结点发送链路的选择

用户态通信机制将网络接口空间映射到了应用层,通信进程可以直接访问本地网络接口获得发送链路的通信状态.MNC 设定数据结构 NICStatus[MAXLINKNUM]记录本结点网络链路的使用状态,由网络接口动态更新.当通信进程发送消息时,直接访问 NICStatus 得到网络链路使用信息,并选择空闲的网络链路发送消息.如果当前没有空闲的网络链路,采用循环选择策略依次选择被占用的网络链路完成发送操作.

(2) 目标结点接收链路的选择

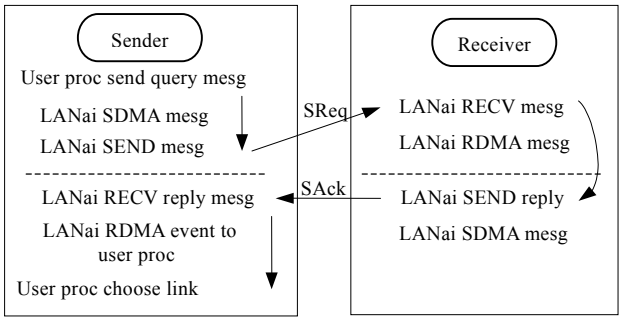


Fig.5 Dynamic link dispatch mechanism
图 5 动态链路分配机制

目标结点接收链路的选择通过源结点通信进程向目标结点发送链路状态查询信息获得.MNC 支持在网络接口/MCP 层对目标结点的网络链路状态进行查询,从而选择接收链路,该过程如图 5 所示.MNC 在 GM Library 中增加了链路状态查询函数,用于发送进程向目标结点发送链路状态查询信息 SReq.SReq 首先被提交给 MCP SDMA 引擎,再由 SEND 引擎发送到目标结点.

当接收方接收到 SReq 消息时,返回包含本结点链路使用状态信息的 SAck 消息.当发送方收到 SAck 消息时,分别经 RECV 和 RDMA 引擎提交到用户层.从而可以选择处于 FREE 状态的接收链路或顺序选择接收链路.

2.4 MNC对长消息通信的处理

短消息的低延迟和长消息的高带宽性能是通信系统的追求目标,而中长消息的延迟性能也非常重要.当有大量控制消息传递或同步操作时,通信延迟性能对计算效率有很大影响.在有多个空闲的网络链路可用时,MNC 将长消息分拆后并行利用多条网络链路发送,以降低通信延迟,同时提高链路利用率.

消息在发送方分拆成消息片发送后,要保证在接收方接收到所有消息片,并按顺序组合成完整的数据包.为了保证所有消息片能尽量同时到达接收方,从而缩短最终消息的发送完成,所有的消息片大小相等,并使用不同的发送和接收链路并行传送.发送进程按照消息片大小依次指定消息片的起始地址和长度发送.而在接收方按同样的方式提供接收缓冲区地址.为了保证消息片在接收方按顺序重新组合,约定接收方的接收缓冲区和发送缓冲区的分块方式相同.图 6 是长消息分拆与组合示意图.

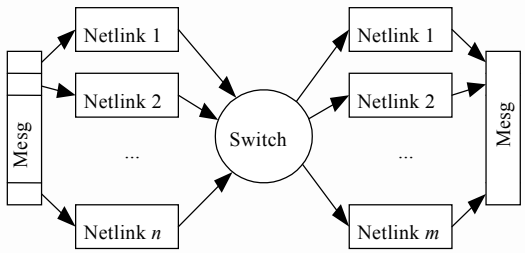
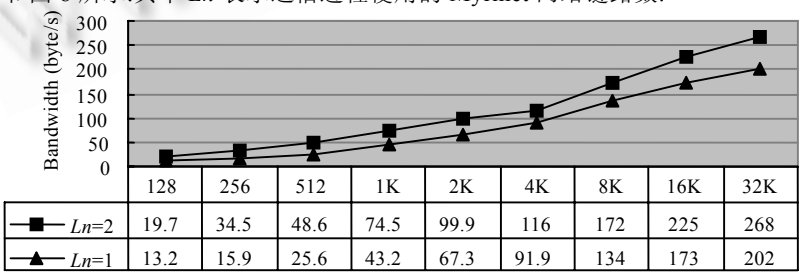


Fig.6 Message segmentation and combination
图 6 消息的分割与组合

将消息均匀分割后,最终的延迟性能将决定于每片短消息的通信延迟加上消息分割和组合而增加的延迟开销.对于短消息而言,链路选择开销相对于消息传输延迟而言可能很大,所以直接利用单链路发送以保证延迟性能.

3 实验结果与分析

本文在测试平台上测试了使用双 Myrinet 链路互连的 MNC 通信性能,并与单链路 GM 通信性能作了比较,分别如图 7 和图 8 所示.其中 *Ln* 表示通信进程使用的 Myrinet 网络链路数.



	128	256	512	1K	2K	4K	8K	16K	32K
Ln=2	19.7	34.5	48.6	74.5	99.9	116	172	225	268
Ln=1	13.2	15.9	25.6	43.2	67.3	91.9	134	173	202

Message size (byte)

Fig.7 Bandwidth
图 7 通信带宽

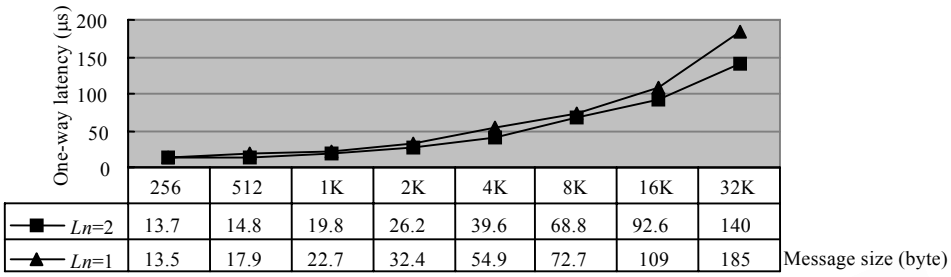


Fig.8 One-Way latency for medium and long messages

图 8 中长消息延迟

如图 7 所示,在使用 1 路和 2 路 Myrinet 互连的情况下,点到点的通信带宽最高分别达到 202MB/s 和 268MB/s,双链路连接比单链路连接通信性能提高了约 34%,已经超过单链路 Myrinet 的硬件带宽(250MB/s).对于计算性能要求不高而通信密集的应用程序,通过在结点上减少计算进程数,增加结点之间网络连接数,可以有效地提高运算效率.图 8 给出了中长消息的延迟性能,多通道通信有效地降低了通信延迟性能.预计随着链路数的增加,延迟效果会更好.但由于消息在发送方分拆和接收方组合,引入了一定的通信开销.另外,GM 实现了高效单链路通信,所以对延迟性能提高有限.

比较在链路连接数改变的情况下,通信进程之间的通信性能可知:尽管多链路并行通信大大增加了结点之间的通信性能,但硬件层的可用物理通信性能还没有得到充分利用,随着网络链路数的增加,通信性能并没有得到相应程度的提高.这是由以下两方面的因素引起的:

(1) 总线因素的影响

由于 Myrinet 是基于 PCI 或 SBUS 总线结构的网络,用户空间和网络空间的数据交换要经过总线进行,所以总线成为通信性能提高的瓶颈.在测试平台上,64bit/66MHz PCI 总线的最大理论带宽为 528 MB/s,实际对总线读操作(即从用户缓冲区到网络接口缓冲区)的带宽为 315 MB/s,写操作(即从网络接口缓冲区到用户缓冲区)的带宽为 372 MB/s,多链路通信性能始终受到总线性能的限制.随着新一代总线标准如 PCI-X 的出现,总线性能得到很大提高.另外,新一代输入/输出(I/O)技术规范——InfiniBand 架构,从根本上消除了总线瓶颈,多链路互连的通信性能预计会有进一步提高.

(2) 链路的分配和调度引入通信开销

除了总线性能影响多通道通信的性能以外,通信链路的分配和调度也引入了额外开销,特别是对于短消息,网络链路的选择引入了较大的通信开销.有必要进一步改进链路的分配和调度算法.另外,将链路的选择放在网络硬件层实现,是降低链路选择开销的重要途径.

4 相关工作

多通道通信技术是一个较新的研究课题.国际上主要有如下几个机构研究多通道通信技术,以满足大规模计算的要求:

美国 Los Alamos 国家实验室和 Compaq 正在合作建造一台每秒 30 万亿次浮点计算能力的超大规模、基于 QsNet 互连网络的 ASCI Q 机群系统.QsNet[2]是 Quadrics 公司生产的一种高带宽、超低延迟的可扩展高性能互连网络.QsNet 网络在硬件层集成了多网络接口处理能力,有效提高了通信性能.

德国 Mannheim 大学计算机结构实验室开发了一种实验性高性能系统域网络 Atoll[7].Atoll 在一块芯片上集成了 4 路独立的网络接口、4 路独立的链路接口和 8*8Crossbar 交换器,从而实现了“芯片上的网络”这样一种结构.Atoll 将多网络链路集成到一块通信芯片内,增加了对 SMP 系统的通信支持,支持多路数据同时通信.

目前,上述工作存在以下不足.(1) 以上系统都采用了专用的新型互连网络,虽然性能较高,但通信费用昂贵,不适合应用于一般计算环境.(2) 在硬件层集成多网络接口对网络接口内置通信处理器的处理能力要求相当高.另外,硬件层实现也不便于该项技术应用于其他网络,灵活性不高.本文基于 Myrinet/GM 的 MNC 实现了在软件层对多路 Myrinet 网络的控制和使用,其设计和实现思想可以应用于其他互连网络,从而增加了灵活性,并提

高了多通道通信的性价比.

5 小 结

本文分析了单链路环境下通信性能难以提高的不足,并设计和实现了支持多路 Myrinet 并行通信的协议层——MNC.MNC 有效地利用了多路 Myrinet 网络物理通信性能,突破了单链路网络互连的性能瓶颈,提高了应用层通信性能.虽然 MNC 基于 Myrinet 网络平台,其设计思想同样适用于其他互连网络.我们正在研究在少数机群计算机结点之间使用多路 Myrinet 网络互连,以增加通信性能、降低反应时间等,适于实时控制或模拟等应用.

References:

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

[2] Petrini F, Feng WC, Hoisie A, Coll S, Frachtenberg E. The quadrics network (QsNet): high-performance clustering technology. In: Proceedings of the 9th IEEE Hot Interconnects (HotI 2001). IEEE Computer Society Press, 2001.125~133

[3] 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 ISCA. Cold Coast: ACM Press, 1992. 256~266.

[4] Prylli KE, Tourancheau B. BIP: a new protocol designed for high-performance networking on Myrinet. In: Proceedings of the International Parallel Processing Symposium 1998. Orlando: IEEE Computer Society Press, 1998. 472~485.

[5] Myricom. The gm API. 1998. http://www.myri.com/GM/doc/gm_toc.html.

[6] Coll S, Frachtenberg E, Petrini F, Hoisie A, Gurvits L. Using multirail networks in high-performance clusters. In:IEEE Cluster 2001. Newpert Beach: IEEE Computer Society Press, 2001. 15~26.

[7] Bruning U, Schalicke L. ATOLL: a high-performance communication device for parallel systems. In: IEEE, ed. Proceedings of the 1997 Conference on Advances in Parallel and Distributed Computing. Shanghai: IEEE Computer Society Press, 1997. 228~234.