

集群高速互连网络分析^{*}

李 涛 陈宇明 赵精龙 倪长顺 杨愚鲁

(南开大学计算机科学与技术系 天津 300071)

摘 要 集群是当今高性能计算领域的重要发展方向,高速互连网络是构建高性能集群系统的关键技术,它是影响集群系统整体性能的关键因素。本文对几种用于集群互连的高带宽、低延迟高速互连网络进行了分析与比较,最后指出了高速互连网络的未来发展。

关键词 集群,高性能计算,互连网络

Analysis on High Speed Interconnection Networks of Cluster System

LI Tao CHEN Yu-Ming ZHAO Jing-Long NI Chang-Shun YANG Yu-Lu

(Department of Computer Science and Technology, Nankai University, Tianjin 300071)

Abstract In the high performance computing area, cluster has been playing an increasingly important role. High-speed interconnection network, which directly affects the whole performance of cluster system, is very critical in building cluster system. A few of high speed interconnection networks are analyzed and compared in this paper, all of which have high bandwidth and low latency. In the end, the interconnection technology that will be developed in the future is pointed out.

Keywords Cluster, High performance computing, Interconnection network

1 引言

在很多科学和工程应用领域中,对计算能力的需求越来越高,这在很长一段时间内将依靠并行处理技术来解决。近年来,随着高性能工作站和网络部件商品化的发展,使 PC 机或工作站网络成为并行处理的理想工具,取代了传统的昂贵而特制的并行超级计算机,转向使用廉价的集群(Cluster)系统为大规模应用提供高性能计算能力。集群是由网络连接在一起的很多工作站或 PC 机构成的并行/分布式处理系统,像一个单独集成的计算资源一样工作,它充分利用了网络和单机性能的优势。它具有投资小,研制周期短,性能价格比高,可扩展性高,使用方便等特点,再加上可移植异构编程环境 PVM 和标准的消息传递平台 MPI 等并行编程环境的日益普及,目前已经成为高性能计算领域重要的发展方向^[1~3]。

拓扑结构是传统并行超级计算机研究中的重要内容,各处理器之间的互连方式是影响系统性能的关键因素,尤其是对于大规模并行系统更是如此。在松散耦合的集群中,节点之间和节点内各部分之间也需要高效的互连方式,互连网络是集群的主要性能瓶颈之一,其容量和性能直接影响了整个系统对高性能计算的适用性,提高集群的整体性能是当今高性能计算领域的研究热点。为此,专门设计开发了高速互连网络硬件,并已经实际应用于集群中。本文对几种应用于集群的高效互连网络进行了分析比较,最后讨论了互连网络的发展趋势。

2 集群对互连网络的要求

并行计算为解决单处理器速度瓶颈的最好方法之一。一

方面,传统的并行计算机研制周期长、价格昂贵、可扩展性不好;另一方面,随着硬件技术的发展,高性能工作站和网络技术的性能不断提高而价格不断下降,导致了低价商品化集群的出现。集群中的各节点都是一个完整的系统,每个节点都有自己完整的操作系统和本地磁盘,可以是工作站,也可以是 PC 机。所有的节点通过互连网络(如以太网)进行连接,网络接口与节点的 I/O 总线松耦合相连。集群成为构建可扩展并行计算机的主要趋势。

为了更好地发挥集群的整体性能,处理越来越复杂的商业服务和科学计算问题,它需要的是一种特殊优化的网络机制和低成本、高性能的互连网络来提高系统的并行性能和高可用性。首先,网络的性能主要由带宽和延迟进行测量,带宽是每秒通过网络可传输的数据量,如果应用程序发送大消息,则系统性能受带宽的影响比较大。延迟是消息从一个节点到达另一个节点的时间,其中还包括创建消息这样的软件开销,如果应用程序需要发送大量的短消息,则系统性能受网络延迟的影响比较大。其次,兼容性是选择网络硬件时的重要参考,网络应该能够与节点硬件和操作系统互操作,能够有效地支持上层所需的通信协议等。再次,共享内存模型比消息传递模型能够支持更小的任务粒度,提高集群的并行度,而且还可以将原来的并行程序比较容易地移植到集群上。因此,随着大规模科学和工程应用对计算能力的需求,集群需要高带宽、低延迟、可扩展性和通用性好的互连网络硬件,甚至可以集群专门制造互连网络硬件来满足其需要。

现有的大多数高效互连网络都属于系统区域网络 SAN。与其他网络相比,SAN 实现了更高的带宽,更复杂的控制数据传输机制,更好的用户级通信,甚至实现了全局地址空间接

^{*} 基金项目:天津市自然科学基金(编号:043600111)资助。李 涛 博士研究生,讲师,主要研究方向为并行与分布计算、高性能网络。杨愚鲁 教授,博士生导师,主要研究方向为并行计算机系统结构、网络计算、可重构系统结构。

口。SAN与局域网LAN(如以太网、Fast-Ethernet等)、广域网WAN(如ATM)一样,已经成功应用于现在的高性能集群中,为集群提供了良好的通信网络环境。

3 高速互连网络

在集群系统中,一个高速互连网络是其性能发挥的重要基础。目前的局域网连接所提供的带宽已经远远不能满足高性能计算的需要,因此很多研究机构和公司为集群特制了高速互连网络,用于建立高性能集群。下面将对几种高效互连网络进行分析与比较。

3.1 概述

高性能并行接口(HiPPI)^[4]是连接异构的主机与外围设备的快速数据移动器,适用于分级互连,保证资源的有效使用。现已制定了与LAN、WAN的互连标准,可将整个系统连接到其他网络上。但HiPPI本身不提供彻底的、通用的集群互连,需要一组不牺牲数据传输速度的多处理器管理系统。

可扩展一致性接口(SCI)^[5]符合IEEE 1596-1992标准,是一个为集群提供高带宽、低延迟的信息共享和信息通信服务的可扩展的系统结构,是综合了处理器、内存、I/O总线和LAN的先进替代物,可以连接普通总线以及I/O的接口。SCI使用带有透明缓存的分布式共享内存,通过隐藏远端数据访问的开销以及消除软件开销而提高性能,对于局域距离或者更短距离来说是一个很有吸引力的互连网络。但是,其可扩展性受现有开关的限制,而且部件比较昂贵,并没有得到广泛应用。

ServerNet-II^[6]系统区域网络在硬件上支持虚拟接口系统结构,它实现了虚拟接口的消息传递和远程内存操作模型。所有的远程内存操作都通过虚地址进行,可以在远程CPU不参与的情况下进行远程存储器读操作。通信协议在本地硬件上实现,消除了通常因消息传递造成的CPU开销,并且允许进程实现通信和计算的重叠而不需要特别的编程技巧。其硬件的高可靠性使实现低开销协议层和驱动软件成为可能,消除了I/O瓶颈,实现了许多对集群计算非常有用的特性。但缺乏通用性,并且价格昂贵,很少被研究者接受。

Myrinet^[7]是全双工包交换新型网络,是专用的、高性能的内部连接,目标是连接商品化部件来构成拓扑结构任意的SAN或LAN集群。它使用自初始化、低延迟的直通式路由开关,主机接口通过自动映射网络配置来提供容错、选择路由以及控制包通信,通信通道提供流量控制、包成帧和错误控制,高效的软件允许用户进程和网络直接进行通信。它具有低延迟、高吞吐量、耐故障和可扩展性好等支持集群计算应用的特点,具有很好的性能价格比,不过,25米的最大传输距离成为系统进一步扩展的最大障碍。

内存通道(MC)^[8]网络是以反射内存(Reflective Memory)^[9]技术为基础的集群专用连接,是一种与SAN完全不同的方式,MC2实现了点到点、全双工、交换式的8×8门闩。内存通道网络通过结点间地址空间映射提供结点间的虚拟共享内存,发送和接收消息的开销与访问本地主存类似,单个数据传输的延迟非常低,数据传输率很高。通过把结点的虚拟地址空间映射到内存通道接口来建立结点间连接,实现了直接的用户级消息传递,使得工业标准通信接口和并行程序设计的效率更高。基于硬件的阻碍应答和快速加锁原语,以及节点失败的检测和隔离,使得它成为理想的集群互连方式。但对集群节点数的限制阻碍了MC的应用。

Synfinity II^[10]系统区域网络提供一个通用的互连基础设施,适于连接微处理器、内存、I/O设备和存储/网络子系统,提供对消息传递和共享内存两种并行编程模型的支持,提供多种多样的数据传输和通信服务,并且许多通信机制都用硬件直接支持,它在时钟周期的两个边缘都用于传输数据,它是达到最高可能物理带宽的有效途径,是集群中消息传递的很好媒介,但可扩展能力差。

光纤通道(FC)^[11]是与SAN紧密相关的技术,在通信设备间提供直接或者交换的点对点连接,真正实现了通道和网络标准的结合,带宽高、可扩展性好。FC由五个功能层次组成,提供三种类型的服务和四种数据传输速率,硬件集中、开销低,消除了传统数据通信方法中的性能瓶颈。它将数据传输从控制协议中分离出来,实现了在单一物理连接上的多协议支持能力。既可用于前端通信网络,又可用于后端存储网络,现在已经成为用于SAN环境的开放标准。但是,其协议复杂,软件开销比较高。

ATOLL(Atomic Low Latency)^[12]是一种新型互连技术,使用ASIC芯片设计方式,集成了一个完整SAN的所有逻辑部件,它是一种将所有网络逻辑紧密的附加在主机上的硬件集成的根本性的设计方法。每个主机端口上都装有硬件实现的DMA引擎以实现消息句柄的功能,其API通过PIO提供了数据传输的软件消息句柄,适于处理短消息,只是没有并发处理功能。ATOLL能够连续接收到达的消息,自动的、以完全并发的方式暂存到主机内存中,通信延迟很低。ATOLL非常满足一个主机连接到四个邻居节点的拓扑,如2D-Mesh和Hypercube等,而且使用额外的网络接口卡很容易地增加每个主机的链路总数。

CLAN^[13]是一种高性能用户级网络,给应用提供了简单的低级别接口,用于数据传输的面向连接的非一致分布式共享内存(DSM)和用于同步的“绊网(tripwire)”。这个接口仅仅使用网络接口卡的硬件实现,可支持很多种应用和通信框架,但不支持传统的DSM通信模型。通过使用标准的处理器写指令就可以实现应用程序间的交互,消息传输开销小,网络延迟非常低,而且实现相对简单,并能有效地支持MPI、VIA、TCP/IP和CORBA等。但是,小数据包消息在目前只用于连接管理。

QsNetII^[14]引入了一些最小化短消息延迟的特征优化处理器间的通信性能,由通信处理器Elan4和交换单元Elite4两个ASIC组成。Elan4能够实现用户空间到用户空间的RDMA操作,并为并行处理所需的高层消息传递协议提供本地处理能力。短事务引擎(STEN)装配短数据包并传送到网络中,优化了短消息的读写操作以及协议控制,并以流水线的方式传送消息,延迟非常低。由Elite4构成基数为4的胖树网络,随着网络规模的扩大,二分带宽能够线性扩展,而且节点间可选择的路由较多,容易实现全局网络操作,适合构建高性能和高可扩展性的集群,但价格比较昂贵。

3.2 高速互连网络分析

各种高速互连网络都提供了高带宽和低延迟功能,最慢的HiPPI的单工单链路为800Mbps,最快的Synfinity II网络的速度最高可达到1.6Gbps,其它网络也都有1Gbps以上的传输速度,而且它们的聚合链路能够达到更高的传输速度。随着激光器件和光波导及其原件制造技术的进步,光互连的成本和体积都在下降,并正在逐步成为各种高性能互连网络的重要选择,当然所有的网络都支持铜缆,只是在传输距离上

远不如光纤,各种网络的传输距离一般都是几十米左右,消息传输延迟很低,满足了用于高性能计算的集群的构建。如果要在更大的范围内建立集群,可以选择传输距离较远的千兆网或者 ATM 网络来进行互连,但这会增加消息传输的延迟时间,降低系统性能。

高速互连网络为上层通信机制和并行计算环境提供了良好的互连基础,虚拟接口系统结构 VIA (Virtual Interface Architecture)^[15]是由 Intel、Compaq 和 Microsoft 提出的一种通信结构的开放工业标准,定义了把应用程序直接连接到网络硬件的快速接口,试图将用户级通信结构标准化。它有效地避免了操作系统在消息发送和接收过程中的干预,同时大大简化了通信协议。其目标是让系统区域网络成为高带宽、低延迟、极低出错率、可扩展以及高可用的互连结构,并允许在任意集群组件与任意其他组件之间以最小处理器间隔进行通信。ServerNet-II、CLAN 能够在硬件上支持 VIA,SCI、Myrinet 借助上层软件也支持 VIA,QsNetII 虽然不支持 VIA,但本身也实现了用户级的消息通信机制,使上层通信性能达到最优,以满足高性能集群的需求。

但是,各个高效互连网络也不是十分完善的,如在所能够连接的节点数目和可扩展能力方面,HiPPI、Synfinity II 和 MC 都不够理想,后者最多只能连接 64 个节点。CLAN、Myrinet 对短消息的支持比较差,FC 的软件开销特别大,将其速度限制到了 255Mbps。另外,大部分网络的价格比较昂贵、通用性差,很难进行市场化。即使这样,现在的互连网络还是保证了 1Gbps 到几 Gbps 的性能容量,光介质的使用增强了距离容量,在一定程度上极大地满足了集群的互连需求。

4 高速互连网络的未来发展

随着计算能力向数据中心的集中,消除性能瓶颈和改进系统管理变得更加重要。需要一种新型互连技术将计算节点、外部海量存储设备以及外部网络作为相互独立的物理单元互连起来,使各物理单元通过中央高速交换网络进行互连,实现从以计算节点为核心的互连转向以计算节点和外部存储节点为核心的互连,从而改变现有系统的整体构造,使系统向模块化的方向发展。现有的千兆以太网、Myrinet、SCI 等互连网络,其带宽和总线的利用率都不能充分满足计算结点互连、计算结点和存储设备互连的需要。

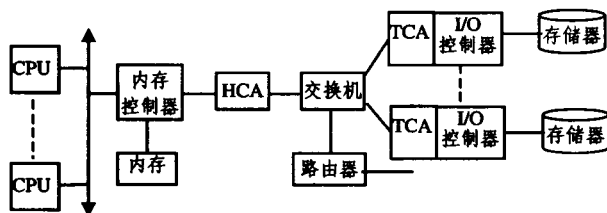


图 1 InfiniBand

InfiniBand^[16]是为此而提出的一种新型的、使用标准的内存-内存互连技术的高性能、高带宽、高扩展性的基于交换的、点对点的高速互连网络。它简化了处理结点、存储结点之间的连接方式,提高了连接带宽,并能与其他外围设备连接,也可在一个节点内部使用 InfiniBand 结构互连多个处理单元。它采用点到点的交换结构提高了系统容错性和可扩展性,实现了单条链路 2.5~30Gbps 的数据传输,传输距离可达数千米,其硬件设计及专用协议都实现了 VIA,而且它可

以和现有的 FC、ServerNet 等互连网络相互操作,这种模块化和可扩展的体系结构满足了集群中的高扩展性和可靠性。

总结 在科学探索和工程实践过程中,问题的规模越来越大,数据量飞速增长,对计算能力的需求也越来越高。虽然集群为大规模应用提供了高性能计算环境,但有效计算能力仍然比较低。使用标准的测试程序对集群进行测量所获得的最大性能,相当一部分达不到其理论峰值的一半,互连技术是首要的制约因素。高性能计算环境对于计算机内部及处理器或计算机之间互连的速率要求越来越高,互连网络成为集群中最关键的技术,为了更好地提供有效计算能力,就需要选择特殊优化的网络机制和低成本、高性能的互连网络来提高集群的并行性能。本文所讨论的现有的高速互连网络都具有 Gbps 级的带宽、极低的延迟和高可靠性的网络硬件,支持用户级网络通信,尤其是 InfiniBand 这种标准的内存-内存互连技术的提出,大大简化了节点之间的连接方式,大大提高了节点之间的连接带宽和连接距离,增强了系统的可靠性和可扩展能力,它是未来互连网络的重要发展方向。总之,高性能的互连网络为集群的整体性能提供了良好的硬件保证,并且大大缩小了集群与大规模并行处理机之间的性能差距。但是,互连网络的价格是制约集群的整体性能价格比的关键因素,为集群选择互连网络时要综合考虑,达到最高的性能价格比。

参考文献

- 1 <http://www.top500.org/>
- 2 Hwang Kai, Xu Zhiwei. Scalable Parallel Computing-Technology, Architecture, Programming[M]. China Machine Press, 1999
- 3 郑纬民,石威,汪东升,等译. 高性能集群计算:结构与系统[M]. 电子工业出版社, 2001
- 4 Tolmie D, Flanagan D. HIPPI: It's Not Just for Supercomputers Anymore[J]. Issue of Data Communications, 1995
- 5 Gustavson D B, Li Qiang. The Scalable Coherent Interface (SCI) [J]. IEEE Communications Magazine, 1996, 34(8): 52~63
- 6 Heirich A, Garcia D, Knowles M, Horst R. ServerNet-II: a Reliable Interconnect for Scalable High Performance Cluster Computing[S]. Compaq Computer Corporation, Tandem Division, 1998
- 7 Boden N J, Cohen D, Felderman R E, et al. Myrinet: A Gigabit-per-Second Local Area Network[J]. IEEE Micro, 1995, 15(1)
- 8 Fillo M, Gillett R B. Architecture and Implementation of MEMORY CHANNEL 2[J]. Digital Technical Journal, 1997, 9(1)
- 9 Jacunski M G, et al. Low-Latency Message Passing for Reflective Memory Networks[A]. In: Workshop on Communication, Architecture and Applications for Network-Based Parallel Computing (CANPC)[C], 1999
- 10 Koyanagi Y, Horie T, Miyoshi T, et al. Synfinity II - a high-speed interconnect with 2 GBytes/sec self-configurable physical link [A]. Hot Interconnects 9[C], 2001. 23~29
- 11 Jurgens C. Fibre Channel: A connection to the future[J]. Computer, 1995, 28(8): 88~90
- 12 Kluge J, Fischer M, Bruning U, et al. ATOLL - A Next Generation System Area Network[A]. The Fourth International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region[C], 2000, 1: 122~127
- 13 Riddoch D, Mansley K, Pope S. Distributed Computing with the CLAN Network[A]. In: Proc. of the 27th Annual IEEE Conf. on Local Computer Networks[C], 2002
- 14 <http://www.quadrics.com/>
- 15 Compaq Computer Corp., Intel Corporation and Microsoft Corporation. Virtual Interface Architecture Specification Version 1.0, 1997
- 16 InfiniBandSM Trade Association, InfiniBandSM Architecture Specification Volume 2 Release 1.1, 2002