MPP 系统的互连通信技术研究*)

Research on the Interconnect Communication Technology in MPP System

刘 燕 杨晓东

(国防科技大学计算机系 长沙 410073)

Abstract Interconnect communication is the crucial factor to influence the performance and efficiency of MPP system. In order to reduce communication latency and increase the network throughput of MPP, in this paper, we study the interconnect network technologies, focus on network interface, communication mode, interconnect network and pipelined-channel technologies.

Keywords Massively parallel processors(MPP), Interconnect communication

从系统设计的角度看,大规模并行处理系统(massively parallel processors,简称 MPP)是将适度并行处理技术中处理机数受限的问题转化成互连网络的通信受限问题。当系统中成千上万个处理结点都频繁地通过互连网络交换信息时,网络中资源冲突和拥塞将严重地限制 MPP 系统整体性能的发挥,且随着硬件技术和工艺的发展、处理速度的不断增快,该问题变得日益突出。为缓解乃至解决 MPP 系统中通信和计算能力的失衡,必须对系统中的互连通信技术进行深入研究,以实现结点间的高速通信。

1 互连通信性能指标

互连通信有两个重要指标:通信带宽(B)和通信延迟(L)。前者指系统在每秒内发送或接收到的消息的字节数,它取决于结点的体系结构和通信机制;而消息在网络中的通信延迟主要由以下三部分组成:

1)建立延迟(Start-up Latency)。源结点的处理器到路由器的发送延迟和目的结点的路由器到处理器的接收延迟之和,主要取决于结点与路由器之间的网络接口的性能、通信方式和通信软件设计,包括通信协议、应用接口、操作系统等。

2)网络延迟(Network Latency)。由源结点的路由器到目的结点的路由器所经过的时间,由互连网络决定,它对通信开销的影响还取决于通信方式。

3)阻塞延迟(Blocking Latency)。由于通道冲突、网络冲突等而导致的阻塞时间,主要取决于互连网络中采用的路由算法和流量控制策略等。

尽可能地提高通信带宽和减少通信延迟是提高 MPP 系统性能和应用效率的关键所在,因此高带宽低延迟的互连通信系统的设计已成为 MPP 系统中一个极为关键的技术,为达此目的,系统需要在结点结构、网络接口、互连网络技术和通信机制等方面进行全面地分析和研究,综合考虑,以设计出整体性能相对最优的互连通信系统。

2 互连通信关键技术

设计互连通信系统的关键在于降低通信延迟、提高网络带宽和带宽的利用率。下面将分析 MPP 互连通信系统中的一些关键技术问题。

(1)**网络接口**

网络接口是处理结点和互连网络进行消息传输的连接部件,传统的网络接口是一个不具备任何处理能力的消息发收部件。由于这种接口不能对消息作任何处理,消息发收的一切加工操作都要由处理器来完成,通信占用了计算时间,因此实际通信效率和处理效率都不高。为提高通信效率,很多 MPP 系统或模型对传统网络接口进行改进,或对传统的工作方式改进、使通信与计算尽可能重叠;或加强网络

*)本课题得到国家 863 计划和九五国防预研基金资助。刘 燕 博士,主要研究方向为高性能计算机体系结构。杨晓东 教授,博士生导师,主要研究领域为高性能计算机体系结构、分布与并行处理。

接口的功能，提供相应的硬件支持以获得最小的软件消息传递开销。

基于DMA的网络接口设计是对传统网络接口的最早改进，在很多MPP系统中得以采用。基于DMA的网络接口设计较为简单，为进一步提高通信效率，很多MPP系统在互连网络设计中加强网络接口的功能，提供更高级的硬件支持，以减少处理器的参与，达到快速通信，典型的如Cray T3D中采用大量的硬件支持，M－Cache模型中接口部件具有消息的组装、接收、缓冲、支持查询等功能，但该种方法内存开销较大，所能支持的通信协议有限，硬件实现较为复杂，且消息的处理仍未摆脱处理器的控制，通信对计算的影响仍然较大。

为减少通信对系统性能的影响，人们对传统的工作方式进行改进，提出了数据流、消息驱动等思想，并进行了相应的网络接口设计，典型的如W. J. Dally在MDP模型设计中将网络接口集成到处理器内部、处理器以消息驱动方式进行操作，但这种方法存在消息驱动的计算量小、消息队列管理复杂和在网络接口、处理机、操作系统及编程方式上都需进行重新设计的缺点，如MDP中要求处理机和软件采用与现有的完全不同的设计，因而在实现和应用上的难度较大。

为提高性能且使实现和应用相对简单，目前很多MPP系统在结点上采用通信代理的设计方法，即在结点上增设专门的通信处理器，将网络接口变成一个相对独立的通信系统。通信代理独立完成传统的由操作系统核心完成的底层通信工作，与路由器协同工作，从而使消息的发送和接收完全独立于用于计算的处理器，实现了计算与通信的重叠，而且为通信软件的设计提供了更强的支持和更大的灵活性，便于用户级通信的实现。该种方法典型的为采用双处理器方式的结点，其中一个用于计算、一个用于通信，根据需要可将两个处理器设计成主从方式或对等方式，二者并行工作，将计算处理器从复杂的通信任务中解放出来。

(2)通信方式

通信方式是指消息的传送机制，主要包括通信机制和通信协议及其执行方式，采用新的通信方式是降低建立延迟的一种有效技术措施。

传统的MPP系统多为多计算机系统，其中通信由操作系统控制完成，需要操作系统的大量参与以提供保护、进行缓存管理和辅助执行消息传递协议。随着网络速度的不断提高，消息的网上传输延迟越来越小，通信的延迟越来越取决于通信软件开销，因此，降低通信处理软件开销对提高并行机的性能至关重要。

传统通信方式的软件开销由如下三部分组成：①操作系统的介入开销：通信过程所需系统调用开销和中断时的切换开销；②通信协议：在简单的点到点通信之上为满足用户需要，进行握手应答、地址检验、缓冲管理等，提供容错的、缓冲透明的通信服务；③编程接口：应用程序接口为支持多种通信方法需进行全面的缓冲管理、保护检查、差错校验等。由此可见，这些开销基本属于缓冲管理和切换开销两类，且都是系统为满足应用需要在硬件基础功能上建立的。为降低这两类开销，目前很多系统在通信方式上采用了多种策略，综合起来有以下几种主要方法：

1)采用新的通信机制。寻求不仅能给用户提供低开销的通信服务而且灵活性较好的高效通信机制，通过直接给用户提供一个贴近硬件功能的低层通信机制，使用户根据需要可在其上建立满足应用要求的不同通信协议，从而达到降低开销且提供一定灵活性的目的。有代表性的如主动消息Active message(AM)[1]。

2)减少内存拷贝。传统的消息传输过程中要经过多次用户空间和核心空间之间的消息拷贝，造成了消息发送和接收过程中的很大开销，也造成了系统通信带宽的较大损失，因此在消息传递中实现真正的“零拷贝”成为MPP通信系统设计中的一大力争目标。真正的“零拷贝”是指消息传送过程不需要任何内存拷贝，可直接从用户空间的缓冲中传出，且网络接口可直接将到达消息直接传入用户级的缓冲中，无需进行中间缓存。要实现真正的“零拷贝”技术上有很大困难，目前很多系统中实现的“零拷贝”实质仍需要一个到网络缓存的中间拷贝，而并非真正的“零拷贝”。

3)实现用户级通信协议。传统的通信协议放在操作系统的核心实现，而操作系统核心空间与用户空间的上下文切换会引起较大的通信延迟。为降低开销，可采用高性能的通信软件将操作系统从通信的关键路径中移走，操作系统仅用于初始化和结束处理等辅助操作，将通信协议放在用户空间实现，使消息的传输在用户级执行，协议直接对网络的硬件设备进行操作，不受操作系统的影响，减少操作系统的时间开销和消息拷贝的次数。与传统的采用操作系统进行进程保护不同，这些系统采用比系统调用更低层的机制用于保护，实现有保护的用户级通信。

目前人们已提出或实现了很多通信机制和用户级通信系统，其中有代表性的有 UC Berkeley 大学的 Eicken 等人提出的主动消息 Active message (AM)[1]，实现的用户级接口 UNet，Princeton 大学的 Blumrich 等人在 Shrimp 项目中实现的虚存映射通信机制(VMMC)[2]，Illinois 大学的 S. Pakin 等人提出的快速消息 FM[3]等。在 MPP 系统设计中如何从系统级进行综合考虑，将通信方式与网络接口结合，设计出硬件支持较好、易于实现且高效的通信方式是当前 MPP 系统通信方式设计中的主要问题。

(3)互连网络技术

互连网络是 MPP 系统的关键部件，其基本功能是在大量的处理结点间传送消息，因而其效率直接影响 MPP 系统的性能。互连网络包括拓扑结构、切换方式、流控策略和路由算法四大要素，其中每一要素都直接影响互连网络的通信性能，对这四方面的不同设计形成了不同的互连网络技术。

·互连网络拓扑结构。它定义了互连网络中处理结点间的互连方式，在很大程度上决定着系统的带宽、延迟、可扩展性和对算法的适应性，因此是影响 MPP 系统性能的主要因素之一。互连网络可分为静态互连网络和动态互连网络两种，静态互连网络中结点之间的连接是固定的，处理结点通过一组点到点的链路相互通信；动态互连网络中结点之间的连接是可变的，各结点间的通信通过开关元件转接。

动态互连网络中结点间不论相邻与否均通过开关进行通信，存在可扩展性较差、硬件实现较困难且开关元件成本昂贵的缺点，而相比之下静态互连网络、尤其是低维的静态网络具有可伸缩性好、实现简单、利于大规模集成的优点，因而大多数 MPP 系统均采用低维的静态网络。

·切换技术。是当消息到达中间结点时，对消息怎样处理，怎样选择输入或输出通道的方法。它决定了如何将消息从输入通道送到输出通道。切换技术经过多次改进目前已趋于统一。早期的多处理器系统中采用的切换技术，均取自于计算机通信网所采用的技术，如存储转发方式(Store-and-Forward)、线路交换方式(Circuit Switch)和虚跨步方式(Virtual-cut-Through)等，这些技术均为第一代 MPP 系统中所采用。自从 1987 年 Dally 提出虫孔路由方式(Wormhole routing)以后，由于其所需缓冲区小，在轻载情况下，当传送消息很长时网络的通信延迟与网络直径无关，因而成为 MPP 系统普遍采用的切换技术，甚至成为第二代 MPP 系统的一个重要标志。但采用虫孔方式，当消息受阻时，消息微片停留在所在结点各通道的缓冲上，占用大量通道资源，当网络负载较重时，易阻塞网络，解决死锁相对复杂。虽然近年来人们又提出了很多新的切换技术，典型的如 P. T. Ganghan 等提出的流水的线路交换 PCS 方式、Shin 等人提出的 Hybrid 方式以及类虚跨步等。但迄今为止，虫孔方式仍是 MPP 系统中切换技术的主流。如何针对存在的问题对虫孔路由进行改进，如采用虚通道技术减少拥塞的发生，对死锁问题进行深入研究采用有效的避免死锁的方法等，是目前 MPP 系统设计中采用较多的方法。

·流量控制。互连网络的流量控制策略决定了如何将网络资源分配给各个消息，并解决消息在传送过程中可能遇到的资源请求冲突，主要包括输入/输出通道选择策略、资源冲突解决策略等。好的流控策略应能尽量避免冲突，且减少网络延迟。为避免出现通信中的活锁或饿死等问题，有效的流控策略还必须将网络资源公平地分配给各个消息报文。

·路由算法。它决定了消息在网络中如何选取路径，其性能对网络效率的发挥起着重要作用，因此对其研究极为重要。一个好的路由算法应有三个特点：低通信延迟、高网络吞吐率和易 VLSI 实现。要达到低延迟和高网络吞吐率，路由算法要避免死锁、活锁和饿死，而且要均匀地传送消息，能自适应地沿最短路径路由消息，有足够的容错能力等；硬件上的易实现性是指路由器实现中所需通道、缓冲区、开关和控制逻辑电路占用芯片面积小、引出腿少，控制简单，实现难度低。在路由算法的设计中除需考虑网络延迟和易实现性外，还需重点解决死锁、活锁和饿死等问题，近年来很多算法又将容错作为一项重要指标，力求在自适应性、性能和容错之间寻求最佳平衡。

人们对路由算法已进行了大量研究，提出了很多算法，但它们或存在代价较大、或存在灵活性不够等缺点，且均存在实现难度较大的问题，目前还未出现一种公认好的自适应路由算法。自适应路由算法是目前 MPP 系统中路由算法设计的热点，但其设计和实现难度较大，考虑不周可能出现死锁，目前仅在少数 MPP 系统(如 CrayT3E[4])中实现。如何在已有算法基础上，以低通信延迟、高通信带宽和易 VLSI 实现为设计目标，设计出可扩展性好、自适应性强且易于工程实现的无死锁完全自适应路由算法是当前路由算法研究的重点。

(4)流水通道技术[5]

传统的MPP系统中互连网络和路由器工作在强同步方式下,需要全局分布时钟且各时钟相位要求严格对准,两相邻结点路由器间消息的传送必须保证在一周期内完成,采用这种强同步方式构造由成百上千个结点组成的MPP系统时,不仅时钟分布和时钟相位对准的难度大,且在不同机柜中的结点间的较长连线限制了网络主频的提高,也限制了系统的可扩展性,而在当今工艺条件下很难进一步减少结点间的互连长度,因此必须研究一种新型的网络互连及路由器技术,以打破这种强同步方式。流水通道即是在这一考虑下提出的一种新的高速互连技术。

在一流水通道互连网络里,路由器间采用源同步技术,采样时钟与被传送数据同时由上一个路由器发出,在一条线上可同时传送多个数据,这使得网络的主频独立于线的长度,与系统中结点间连线长度无关,从而有效地提高了网络传输速度。流水通道的思想在广域网与局域网中早已采用,而在多处理机、多计算机的互连中迄今为止还采用较少,特别是在MPP系统中,目前仅有CRAY T3E等少数几个系统中采用了这类互连技术,在采用流水通道的路由器中必须采用特殊技术,如锁相技术,来解决消息的源同步传送和与源同步接收问题。在MPP系统中实现流水通道是大幅度提高网络的传输速率和系统性能的关键技术。

小结 MPP系统中互连通信性能是影响系统性能的主要因素之一,也是决定其使用效率的关键,要实现高性能的互连通信须从网络接口、通信方式、互连网络技术和线上传输技术等多方面协同考虑,以设计出高效的互连通信系统[6]。此外,在MPP系统的设计中,还可采用隐蔽延迟的技术,将上述多种技术综合应用可更好地提高系统的通信性能。

参考文献

1 Eicken T V, et al. Active messages: a mechanism for integrated communication and computation. In: Proc. 19th Int. Symp. Computer Architecture. 1992. 256～266

2 Blumrich M A, et al. Virtual-memory-mapped network interfaces. IEEE Micro, 1995(Feb): 21～28

3 Pakın S, et al. Fast messages (FM): efficient, portable communication for workstation clusters and massively-parallel processors. available at: achien @ cs. uiuc. edu. 1997

4 Scott S L, Thorson G M. The CRAY T3E network: adaptive routing in a high performance 3D torus. available at: sls@cray. com, 1997

5 刘燕,徐炜遐,杨晓东. 流水通道——一种高速MPP系统互连. 计算机学报,1999

6 刘燕. 大规模并行处理系统高速互连通信技术的研究:[博士论文]. 长沙:国防科技大学,1998

(上接第47页)

特点和要求,综合传输效率和可靠性这两个指标进行考虑;

·考虑到大型dVEs系统规模(包括地理跨度和实体数量)的庞大,已不可能,也没有必要向每个用户发送刷新消息。这里需对消息传送的目的地进行选择和过滤。为此,DIS采用了AOI与多信道广播相结合的方法,并在节省带宽方面取得了显著的成效。

DIS是迄今为止最成功的分布式虚拟现实设计标准,在我们研究大型dVEs系统的消息传送机制时,可从DIS得到许多启发,如:标准的消息格式、全分布式模型、呆推断、使用多信道广播过滤刷新消息的尝试等。然而,这并不意味着我们可以简单地使用DIS。DIS是针对一特定应用领域而设计的,因而存在一些局限性。如:扩充性不好,解释其位模式的计算开销较大,是一种个体较"大"的标准。下一代的DIS应具备:更简捷、开放、可扩充以及可动态修改的特点。这种可动态调整的协议对于测试和评价分布式实体的交互的效率是必需的。

参考文献

1 Stytz M R. Distributed Virtual Environments. IEEE Computer Graphics and Application, 1996(May): 19～31

2 Brutzman D. Graphics Internetworking: Bottlenecks and Breakthroughs. Available at: Http: // www. stl. nps. navy. mil/～ brutzman/VRml/breakthroughs. html

3 Macedonia M R, et al. Exploiting Reality with Multicast Groups: A Network Architecture for Large-Scale Virtual Environments. macedonia@cs. nps. navy

4 JDS Pugh. Eliminating Network Traffic Caused by Deterministic Objects in Multi-user Virtual Worlds. jds @postoffice. utas. edu. cn

5 Macedonia M R, et al. NPSNET: A Network Architechture for Large Scale Virtual Environment. macedonia@cs. nps. navy

6 Macedonia M R, et al. A Taxonomy for Networked Virtual Environments. mmacedon@crcg. edu