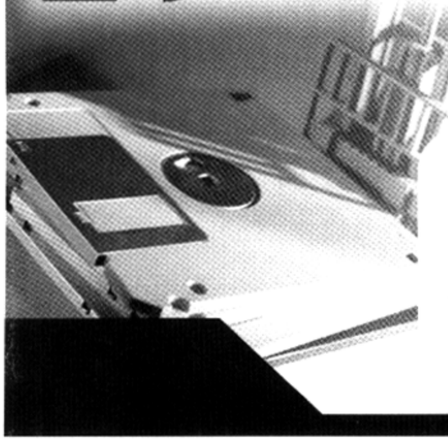
基于InfiniBand网络存储的研究与设计



林永明　常致全（成都四川大学计算机学院　610065）

摘要：本文介绍了传统的存储体系结构的缺陷以及网络存储的现状，并探讨一种新型I/O技术--InfiniBand，进而阐述基于InfiniBand的I/O体系结构的研究，以及IB模拟器的设计。

关键词：InfiniBand　存储局域网　主机适配器　目标适配器　Verbs

1　引言

随着Internet/Intranet以及其他网络相关的各种应用迅速发展，网络上的信息资源呈现出爆炸性增长的趋势，大量信息的处理及传输对信息存储系统提出了空前的要求。传统的附属网络服务器的存储体系中存储设备是通过IDE/SCSI等I/O总线与服务器相连，客户机的数据访问必须通过服务器，随着客户连接数的增加I/O总线将会成为一个潜在的瓶颈。同时服务器能够访问的数据量受制于总线所支持的磁盘的个数，这就限制了单个服务器的容量。这种分布式服务器依靠数据传输网络来执行备份和恢复操作，这些操作会耗尽带宽，使正常的网络传输慢如蜗牛爬。

现有的互连技术已经跟不上计算机的发展，高端的计算机概念如群集，故障安全等要求更好的性能来在处理器结点之间或者处理器结点和IO设备之间传输数据，计算能力越来越向数据中心集中，这些趋势要求更高的带宽和更低的延迟，现有的基于PCI总线的存储结构已不能适应来自应用的越来越高的要求，消除性能瓶颈和改进系统管理变的比以往更加至关重要，I/O子系统是造成很多这类问题的根源。因此，探索新的存储体系结构就非常必要。近年来网络存储成为国际上比较热门的一个研究方向。

2　网络存储

基于传统的附属服务器存储的缺点，在网络存储的存储结构中，存储系统不再通过I/O总线附属于某个特定的服务器或客户机，而是直接通过网络接口与网络直接相连。与传统的附属于服务器的存储系统相比，网络存储系统具有非常好的可扩展性，高带宽，低延迟，而且容易实现集中管理。目前最流行的两种网络存储方案：存储局域网(Storage Area Network)和InfiniBand。存储局域网(SAN)就是一个由服务器和存储设备组成的高速网络，它的用途是在不影响局域网和广域网带宽的情况下实现服务器与存储设备之间的大流量数据传输。本文将着重探讨基于InfiniBand的网络存储。

InfiniBand是“下一代I/O”(NGIO)体系结构和“未来I/O”(FIO)相结合的产物。InfiniBand是集合了整个业界的努力而开发出来的能够替代PCI总线的新标准。InfiniBand是由Intel、Microsoft、Compaq、Dell、IBM、HP以及Sun Microsysterms七家公司共同组成的InfiniBand Trade Association负责研发和执行，这个协会目前至少有120个以上的成员公司。

3　InfiniBand

InfiniBand是一种新型的总线结构，它可以消除目前阻碍服务器和存储系统的瓶颈问题，是一种将服务器、网络设备和存储设备连接在一起的交换结构的I/O技术。它的目标是让计算机甩开总线，通过I/O与CPU的隔离，达到高的带宽和低的延迟。

3.1　InfiniBand结构

IBA(InfiniBand Architecture)定义了依赖于Switch和Router级联的通信结构，可以连接多个独立的处理器平台、I/O平台和I/O设备的局域网，见图1。IBA既可以支持小服务器(一个处理器和很少的I/O设备)，也支持巨大超级计算机(成百的处理器和成千的I/O设备)，并可以自然利用IP协议通过Internet、Intranet和远程计算机系统连接。它提供了一个高带宽，低延迟，安全的，远程的管理环境。

这种体系结构独立于主机操作系统(OS)和处理器平台，它集合了通信和管理，不仅支持I/O而且支持进程间

通信(IPC)。图1中的结点(Node)既可以是处理器结点，也可以是各种存储单元，如磁盘，磁带，RAID甚至可以是SAN。

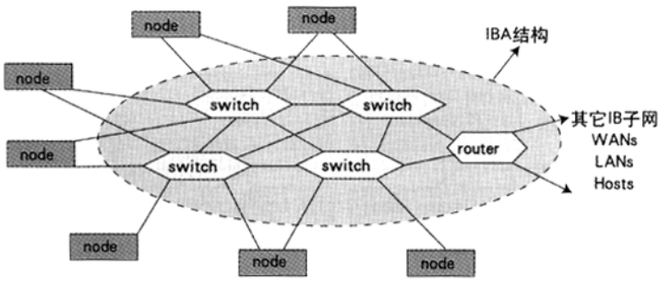


图 1 IBA 系统局域网

InfiniBand 包括四种硬件：

· 主机通道适配器(HCA):将处理器平台连接到IBA 上；

· 目标通道适配器(TCA):将I/O单元连接到IBA 上；

· 交换机(Switch):连接所有的设备；

· 路由器(Router):负责子网间连接。

InfiniBand 主要用来代替服务器中的外设部件互连(PCI)总线。基于InfiniBand的体系结构剔除了PCI总线，处理器平台上内存控制器通过HCA直接连到IBA 上，另一端可能是插有HCA的处理器平台，也可能是插有TCA的I/O单元，其中基于信道的InfiniBand交换器提供点到点的连接。这种新型的体系结构主要特点有：

· 智能的通道适配器(HCA和TCA)能处理所有的I/O操作，释放了CPU 负载；

· 数据传输率高，使用光纤连接可以处理500MB/s到6GB/s的传输速率；

· 增强了扩展能力，基于Switch和Router可无限扩展体系结构，而且链接线缆最长传输距离可以达到10km；

· 具有良好的数据完整性，可用性和可靠性，可以很方便地实现诸如磁盘冗余，关键数据备份，远程群集，远程镜像；

· IBA 提供一种集合所有的通信类型(如SCSI,FC,SAN)的结构，简化了管理的实现；

· 采用IPv6编址，能和目前Internet/Intranet建立高效的连接；

· IBA是一个开放的标准，可以从许多产商那里得到IBA 组件。

3.2 IBA 软件传输接口与 Verbs

IBA 的软件传输定义了通道接口(Channel Interface)的功能和行为，并为用户提供通道。该接口的实现将HCA，与之相关固件和Host软件结合起来。

CI=HCA Hardware+HCA Device Driver

Verbs是InfiniBand引入的一个新概念，Verbs是HCA功能的抽象描述，即指出了CI的功能性(操作性)，见图2。它不是API、DPI，也不是硬件抽象层HAL。但它将影响/指导硬件、设备驱动程序接口、内核编程接口(KPIs)及API的设计。提出Verbs是出于灵活性考虑。IBA并没有要求统一、一致的API或DPI，这就为不同的OSV提供更广阔的设计空间。当前的接口是应用层上的接口，而Verbs只不过是概念层上的接口。OSV要基于这些Verbs向用户提供API、DPI。

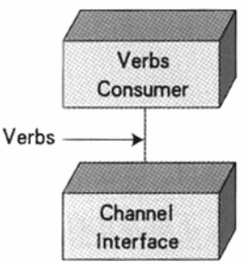


图 2 Verbs

IBA 是基于消息传递的通信机制，该通信机制采用一种灵活，强大的排队模型，见图3。这种通信模型主要包括三个组件：

· 工作请求WR:WR用来向CI递交工作单元，用户只能利用WRs这一机制在工作队列(SQ/RQ)中产生工作，WR也只能用来传递从用户到CI的操作；

· 队列对QPs:QPs分为发送队列SQ和接收队列RQ，QR是硬件系统提供给IBA用户的虚拟接口，它为用户提供了一个虚拟通信端口；

· 完成队列CQ:作为工作请求WR完成的通知机制。

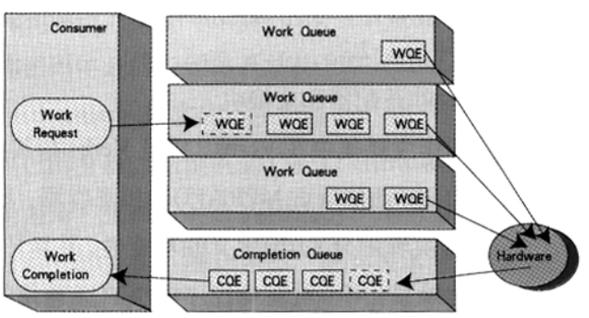


图3 IBA 排队模型

用户的工作操作通过Verbs以WRs的方式递交给CI，WRs在相应的工作队列中产生一个工作队列元素WQE。

CI顺序的执行工作队列中的WQEs。当CI完成一个WQE在以之相关的完成队列CQ中要产生一个工作完成元素CQE，它指明有关工作请求WR的完成信息，这些信息通过异步通知机制返回给用户。我们不难从图2看出，IBA排队通信模型实际上就是WR处理模型：WR如何递交给CI，CI如何处理WR以及WR完成结果如何返回给用户。

从通信的角度来看，在发送队列SQ中CI解释工作队列中的WQEs并产生相应的请求消息，如果有需要的话，QP将消息分片成多个包从正确的端口发送出去。当目的端接收到一个包时，其端口要验证包的完整性。然后CI将接收到的包传递给正确的QP，使用QP上下文关系处理该包并执行必要的操作。如果需要的话，CI会返回一个确认消息。

所以从整体上来看，IBA操作可以分为几个层次：物理层、链路层、网络层、传输层和高层协议。每层上的协议是彼此独立的，下层为它的上层提供服务。

·物理层：构造有效包格式的信号协议；链路层：流控和子网内路由；网络层：子网间路由；传输层：保证包正确的发送，接收和处理，支持可靠连接，不可靠连接，可靠数据包和不可靠数据包等服务；

·应用层：支持SCSI和IP协议，定义基于消息的管理协议。

4 InfiniBand模拟器的设计

由于目前我们无法得到InfiniBand的产品，如HCA，TCA和Switch，所以我们要在现有的设备和环境下模拟InfiniBand的I/O体系结构，我们将遵循InfiniBand标准模拟HCA和TCA，为不久将来全面地实现基于IndiniBand的I/O体系定义、设计出移植性强的软件模块。我们的开发环境：

· Linux平台—2.2

· Myrinet的适配器和交换机

我们的工作主要集中在Verbs，以及Verbs与其他I/O模块(DPI)和IPC通信层如MPI(API)的相互作用。根据InfiniBand标准实现部分Verbs定义的功能性，并用Verbs模拟HCA的传输层，用主机驱动程序模块TCA通用的功能，见图4。

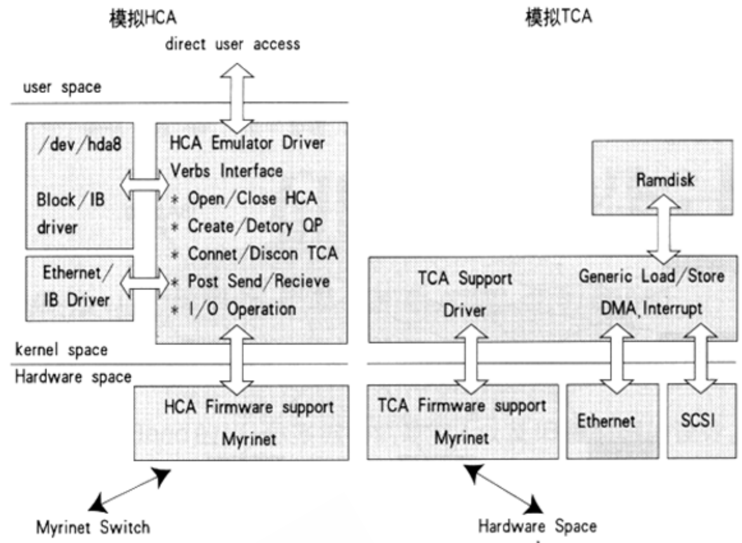


图 4　nfiniBand模拟器的设计结构

在模拟的IBA结构中是以QP作为通信双方的端点，包的寻址是根据目的结点ID+目的结点上QP号。所以实现的两种IBA服务(可靠连接服务和不可靠数据报服务)都是针对目的和源QP而言的。其中可靠连接服务通过CRC、确认和包编号机制来保证消息至多发送一次，且有序无错，包的有序性要同时满足发送有序，执行有序，响应有序，完成有序。建立可靠连接的源和目的结点的QP是一一对应的，并提供Send,Receive,RDMA Send和RDMA Receive等传输语义。不可靠数据报服务只允许发送单包消息，建立不可靠数据报的源和目的结点的QP不必绑定在一起，它们是一对多的关系，只提供Send和Receive传输语义。

在SCSI同IBA结合中，我们采用SRP协议(SCSI RDMA Protocol)，SRP定义了SCSI到IBA的映射关系，将SCSI结构扩展到了IBA。存储驱动程序的操作模型是，主机向目的方发送封装SCSI命令的消息，目的方执行命令并准备传输数据(目的方不向主机方暴露自己的内存)，同时目的方发送状态消息。主机完成请求并调用完成事件句柄。

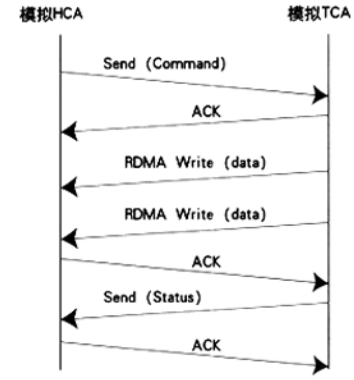


图 5　对磁盘读操作过程

(下转第29页)

(上接第25页)

在基于我们设计的IO体系结构上进行I/O操作的过程可以从图5中清楚地看出。图5描述了Host对磁盘进行一次读操作的过程，通信双方采用可靠连接服务，首先SCSI命令经Host上模拟HCA以消息包的格式发送出去。目标TCA收到消息包对此包确认，同时模拟TCA根据接收的SCSI命令准备以远程写(RDMA Write)的方式将数据传输给Host。Host完成磁盘传来数据的接收相应也返回模拟TCA一个确认。最后TCA要向Host发送一个状态消息，指出这次读操作完成状态，Host对磁盘的写操作与读操作类似，不同的是模拟TCA采用远程读(RDMA Read)方式将Host要写的数据读到磁盘上。

5 结束语

本文介绍了一种新型的I/O技术—InfiniBand，探讨了它的结构和软件传输接口。分析了根据InfiniBand标准HCA和TCA模拟器的设计及工作模式，进而较好的实现基于IBA的I/O体系结构。由于缺乏IBA硬件的支持，所以整个体系结构管理系统的完善，以及存储性能的提高都是需要进一步研究的课题。■

参考文献

1 InfiniBand Architecture Specification Volume 1. Available at http://www.infinibandta.org。

2 InfiniBand Architecture Specification Volume 2. Available at http://www.infinibandta.org。

3 Joe Pelissier. InfiniBand Architectuer Overview. Intel Corporation, available at ftp://download.intel.com/design/servers/future_server_io/documents/sers04%20ss_Pelissier.pdf。

4 Frank L. Berry. InfiniBand Queuing Concepts for Driver Design. Intel Corporation, available at http://www.intel.com/presatations.htm。

5 Paul Grun. Transport Layer Common Functions. InfiniBand Trade Association, available at http://www.infinibandta.org。

6 SCSI RDMA Protocol. American National Standard for Information Systems—Information Technology, available at ftp://ftp.t10.org/t10/drafts/srp/srp-r04.pdf。

7 Bill Bostic. InfiniBand Architecture and Communications. Intel Corporation, available at http://www.intel.com/presatations.htm。