

IPoIB 体系结构及其应用^{*}

刘爱华 钱德沛 董小社 乔楠

(西安交通大学电信学院新型机研究所 西安710049)

IPoIB Architecture and its Application

LIU Ai-Hua QIAN De-Pei DONG Xiao-She QIAO Nan

(Xi'an Jiaotong University, Xi'an, China, 710049)

Abstract InfiniBand (IB) is a high speed, channel based interconnecting technology between systems and devices. It will remove the bottleneck limiting the performance of current servers and storage devices. This article proposes an IP over InfiniBand (IPoIB) architecture and its application. It will function as a bridge between IP and IB architecture in systems IP and IB coexists. It enables the applications to take advantages of IB without changes in upper layers IP and other protocols. We also introduce the IPoIB protocol and show some experiment results in this paper. The experiment results show that IPoIB is suitable to be used in the front-end servers and interface devices for the InfiniBand-based network servers.

Keywords IPoIB, InfiniBand, High-performance servers

1 引言

传统的以太网架构,虽然已经被改进和增强以满足数据中心、高性能服务器的需要,但是由于它本身的带宽限制以及TCP/IP协议的高复杂性,产生带宽瓶颈,消耗服务器的大量CPU和内存资源。这样不仅服务器的带宽成为瓶颈,而且服务器的相当一部分资源被消耗在网络I/O处理而不是应用处理上,并使得它很难扩展^[1]。

InfiniBand体系结构(IBA),作为一种高性能系统互连技术,定义了从物理层到应用层的完整的协议栈和管理机制。它采用RDMA进行通信,OS旁路,高带宽,低延迟,而且具有很好的容错性能,能解决传统以太网的上述不足,消除网络瓶颈^[2]。

但是,InfiniBand的广泛应用,需要一个长期的过程。新型的IB架构将和传统的IP架构共存,并逐步获得广泛的应用。IPoIB(IP over InfiniBand)则为IB与IP的共存提供了一种解决方案。IPoIB允许上层采用传统的IP协议,而下层则采用新型的IB架构,从而现有的应用不做任何改变就能获得IB架构带来的带宽和性能的优势。

2 InfiniBand 体系结构

2.1 InfiniBand 体系结构概述

InfiniBand体系结构规范1.0由InfiniBand行业协会(IBTA)于2000年7月发布。IBTA的目的是制定一个基于通道和交换技术的I/O规范,以适应数据中心、高性能服务器的RAS(Reliability, Availability, Scalability)和性能需求。IBTA于2001年6月又发布了IBA规范1.0.a,对1.0进行了修正。

InfiniBand体系结构(IBA)定义了一个互连多个独立的处理器平台、I/O平台和I/O设备的系统域网络(SAN),它是支持一个或多个计算机系统中并行I/O以及处理器间的通信和管理的平台^[3]。

IBA系统由IB链路互连的处理器节点和I/O单元组成,

其架构包括层叠的交换机和路由器(连接IB子网)。I/O单元可以是单个ASIC IBA附属设备,比如LAN适配器,也可以是大的RAID子系统。

各个IB节点可以通过一台或多台交换机互连,或者直接互连。IB单元通过通道适配器(CA)连接到IB链路上。每个单元支持多个CA,每个CA能提供一个或者多个端口。一个CA相当于一个节点。CA的端口是数据发送的目的地。但是,每个端口可以包括多个队列对(Queue Pairs),它们可以直接被远程寻址。从这个角度说,队列端口号QPN(QP Number)是地址的一部分。InfiniBand拓扑结构如图1所示。

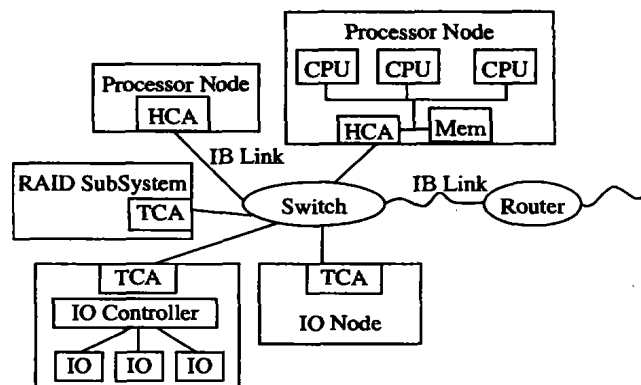


图1 InfiniBand 拓扑结构示意图

2.2 IBA 层次结构

IBA的层次结构自下而上可分为:物理层、链路层、网络层、传输层、Verbs层。在主机节点上层,还有应用和管理层。IBA的层次结构图参见图2。

2.3 IBA 支持的传输模式

IBA支持5种数据传输模式:

1. 不可靠数据报服务(Unreliable Datagram)(无应答-无连接);
2. 可靠数据报服务(Reliable Datagram)(应答-复用);

^{*} 863课题支持,课题编号:2001AA111120。刘爱华 硕士研究生,主要研究计算机系统结构。钱德沛 教授,博士生导师,主要研究高性能计算和计算机网络。董小社 副教授,主要研究机群与高性能计算。乔楠 硕士研究生,主要研究计算机系统结构。

3. 可靠连接服务 (Reliable Connected) (应答-面向连接);

4. 不可靠连接服务 (Unreliable Connected) (无应答-面

向连接);

5. 原始数据报服务 (Raw Datagram) (无应答-无连接)。

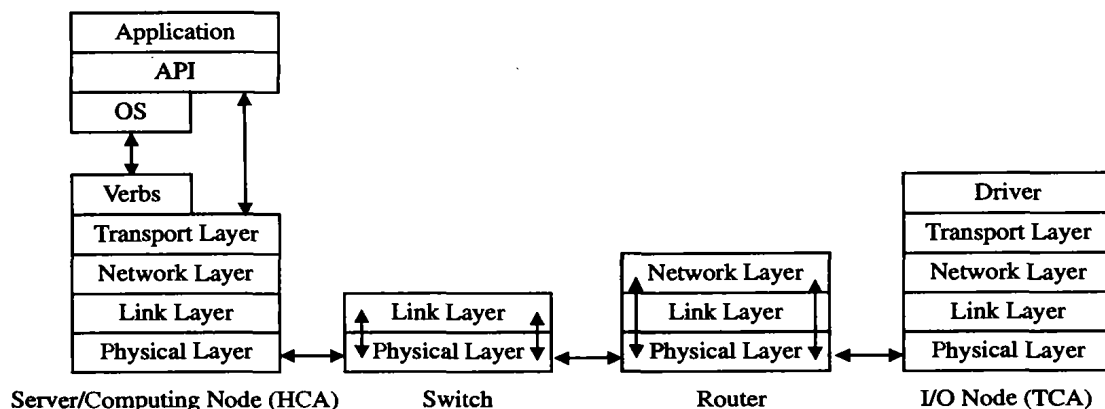


图2 IBA 层次结构图

3 IPoIB 体系结构

IPoIB (IP over InfiniBand) 采用 IB 架构作为链路层, 其上层运行传统的 IP 协议。它的目标是 IP 及其上层协议和应

用不必做任何改变而可以运行在 IB 体系结构之上。IPoIB 采用封装的方法实现 IP 和 IB 报文的转换。即将 IP 报文加上 IB 包头封装为 IB 报文, 从而在 IB 层上进行传输。IPoIB 体系结构如图3所示^[4]。

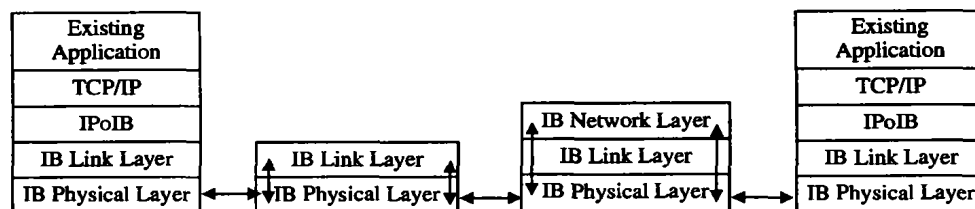


图3 IPoIB 体系结构示意图

由上图可以看出, IPoIB 需要4个部分的功能:

1. 地址解析功能。因 IPoIB 上层为 IP 协议, 使用 IP 网络地址, 而下层为 IB 结构, 使用 IB 地址, 所以 IPoIB 必须实现 IP 地址和 IB 地址之间的转换, 如同传统的 ARP/RARP 的功能一样。

2. 报文封装功能。IPoIB 要将 IPv4/IPv6 报文封装为 IB 帧格式, 以便在 IB 网络中进行传输。

3. 对 IP 多播的支持。

4. 对 IB 链路管理的支持, 提供通道适配器, 交换机等 MIBs 树的管理信息。

4 IPoIB 的实现考虑

在第2节我们已经描述了 IBA 的数据传输模式, IPoIB 可以在其任何一种模式上实现。但是由于不可靠数据报 UD 是 IBA 规范要求所有的 IB 节点都必须支持的传输模式, 它也支持多播、分区和32位的 CRC, 而且从实现的难易程度来说, 不

可靠数据报 UD 模式也是很容易实现的, 因此 IPoIB 的实现采用 IB 的不可靠数据报 UD 模式^[5]。

4.1 帧的格式

所有 IB 上的 IP 和 ARP 传输报文都加了4字节的前缀, 构成 IB 帧的负载, 4字节的前缀的格式如下所示:

16位的类型[15-0]	16位, 保留
--------------	---------

其中, 16位类型字段指示报文的封装类型的值如下表所示:

类型	协议
0x800	IPv4
0x806	ARP
0x8035	RARP
0x86DD	IPv6

封装后, 帧的格式如下:

IB Frame Headers				Payload	IB Trailers	
Local Routing Header	GRH* Header	Base Transport Header	Datagram Extended Transport Header	4-byte Header + IP/ARP	Invariant CRC	Variant CRC

4.2 最大传输单元(MTU)

由于每个 IP/ARP 数据报都包含了上面描述的4字节的

前缀,因此 IP 的 MTU 就是 IB 的 MTU 减去4,例如下表所示:

IPoIB 链路 MTU	IP MTU
4096	4092
2048	2044
1504	1500
1284	1280

因为 IPv4 的 MTU 应不低于1500字节,这样,要求 IPoIB 子网中的组件的支持的最小 MTU 为2048字节,相似地,对 IPv6 也为2048字节。

4.3 硬件地址

IPoIB IP/ARP 的硬件地址格式如下:

10	8	32
保留	QPN[23-0]	
	GID[127-96]	
	GID[95-64]	
	GID[63-32]	
	GID[31-0]	

10	8	16	32
Hardware type		Protocol	
Length of hardware address	Length of protocol address	ARP operation	

其它字段就是发送方和接收方的硬件地址和协议地址:

[Sender hardware address][Sender protocol address]

[Target hardware address][Target protocol address]

这里的硬件地址的长度是20。

5 IPoIB 的应用

IPoIB 可以作为 IP 和 IB 之间的桥梁,在网络服务器特别是集群服务器中有着特殊的作用。IPoIB 可以工作在集群服务器的前端服务器或者是接口机里,而后端的服务器群或者是服务器集群之间通过 IBA 互连。这样,普通的 IP 应用和请求可以通过接口机的 IPoIB 转换为 IB 报文发送到服务器集群中,经过处理后将结果返回。这样,既可以提高集群服务器的性能,又不需要对传统的应用做任何的改动。IPoIB 的应用如图4所示。

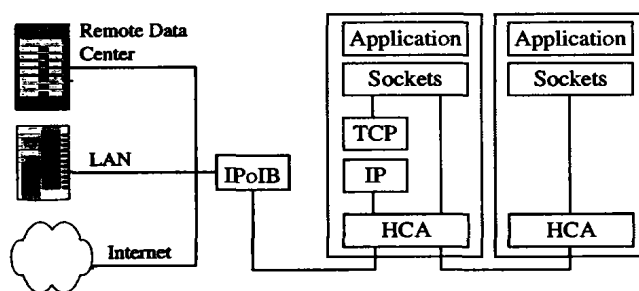


图4 IPoIB 应用示意图

6 IPoIB 的测试与分析

6.1 测试目的

实验的目的是测试 IPoIB 协议以及 VI^[4]协议数据传输的各个方面的性能,主要包括传输速率、IO 响应时间、CPU 利用率等指标,并对二者的性能进行比较。

6.2 测试环境

其中:

1. 8位保留位留作将来使用;

2. 因为 IBA 的节点间通信需要定向到指定的队列端口 QP。这样 IPoIB 的 IP 通信也需要定向到 QPN,所以 ARP 的回应报文也需要解析到 QPN。

3. GID 是端口号分配的全局表示符。

4.4 地址映射

IPv4 的地址解析是通过地址解析协议 ARP[RFC0826]^[6]实现的,而 IPv6 则是通过邻居发现协议实现的,因此,IPoIB 要实现自己的 ARP 协议。这里主要分析 IPv4 的地址解析,IPv6 的地址解析和 IB 规范是类似的^[7]。

4.4.1 IPv4 子网中的地址解析 ARP 报文的报头是在 RFC0826 中定义的。硬件类型是 IANA 规定的,被设置成 32 (十进制)。它的字段在 RFC0826 中定义如下:

我们搭建了一个基于 IBA 与 TCP 环境的试验平台,它的结构如下:

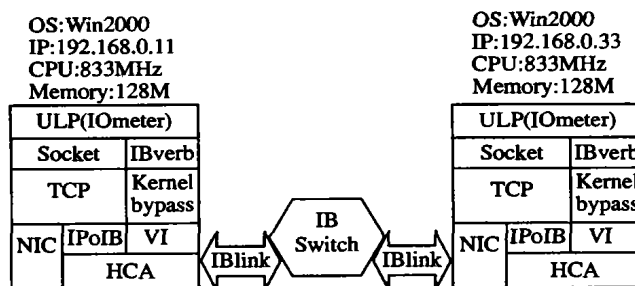


图5 IPoIB 实验示意图

我们使用了 IOMeter 这个软件,它可以模拟不同传输协议下数据传输时的性能指标,并且它直接支持 VI 接口。

6.3 测试结果

当我们使用不同的传输数据量时,IPoIB 以及 VI 所反映的各个方面的性能也不一样,如下图组所示:

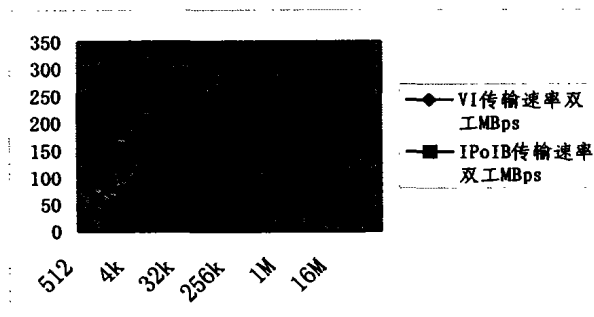


图6 VI 与 IPoIB 的传输速率比较

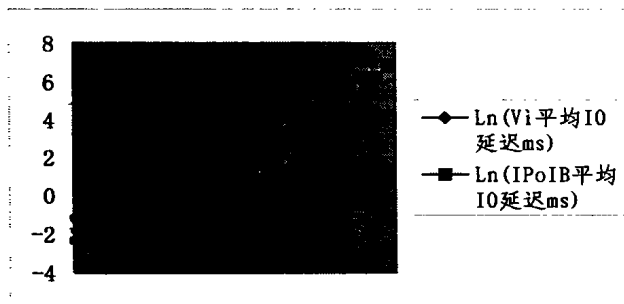


图7 VI与IPoIB的IO响应时间比较

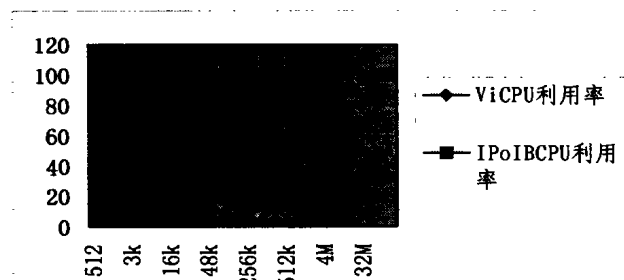


图8 VI与IPoIB的CPU利用率比较

6.4 测试分析

从中,我们可以看出,IPoIB在小尺寸与中等尺寸数据块传输时,传输速率性能上有优势。但是由于CPU在小尺寸传输的情况下负荷很重,同时,在中尺寸数据块传输时的IO响应性能与VI的IO响应性能最接近,所以IPoIB在中尺寸数据块传输时的综合性能最佳。而对于VI而言,由于其基于消息传递的传输机制,以及零拷贝等特性,性能往往到了1M以上的传输量时才显现出优势,并且优势集中体现在CPU的低

负荷上。

可见,对于高性能机群的Web服务以及Email等服务而言,由于它们要求时延低,吞吐量不是很大,短包较多,因此IPoIB是很好的选择。而VI在大数据量传输时可以显示它的优势。

结束语 InfiniBand作为新型的I/O结构,可以缓解传统以太网的带宽瓶颈。IPoIB在新旧技术并存时,可以作为联系IP和IB之间的桥梁,它不需要IP及其以上的协议和应用做任何的改变而利用IB的优点,因此,适于在采用InfiniBand技术的网络服务器的前端服务器或者接口机中使用。

参考文献

- 1 Break Through the Bottleneck. 2002, Voltaire. <http://www.voltaire.com>
- 2 InfiniBand™ Architecture Specification Volume 1, Release 1.0, 2000, InfiniBand Trade Association. <http://www.infinibandta.org>
- 3 InfiniBand™ Architecture Specification Volume 1, Release 1.0.a, 2001, InfiniBand Trade Association. <http://www.infinibandta.org>
- 4 draft-ietf-ipoib-architecture-01. txt, Vivek Kashyap. <http://www.ietf.org>
- 5 draft-ietf-ipoib-ip-over-infiniband-01. txt, Vivek Kashyap, H. K. Jerry Chu. <http://www.ietf.org>
- 6 rfc826-Ethernet Address Resolution Protocol. txt, David C. Plummer. <http://www.ietf.org>
- 7 rfc2461-Neighbor Discovery for IP Version 6 (IPv6). txt, T. Narten, E. Nordmark, W. Simpson, Daydreamer. <http://www.ietf.org>
- 8 Virtual Interface Architecture Specification. <http://www.vidf.org>

(上接第80页)

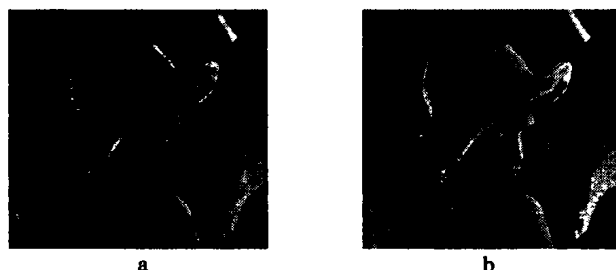


图4 a为图3的恢复图像, b为标准Lena图像

之所以选取图像2而不是图像3的权值矩阵作为恢复图像3的参数,是因为权值矩阵的作用是为了去除坏像素和降低振铃效应,这样做可以保证图像3中增强了的高频信息不至于被当作坏像素而被屏蔽。从视觉效果来看,本文的方法是可行的。本文没有比较模糊图像恢复前后的信噪比,因为信噪比并不是一个好的评价标准,它反映的也只是图像的统计特征,使用信噪比对重构图像的评价与人的主观视觉评价并不一致。

本文考虑的模糊图像并没有包含噪声,因此在使用本文方法之前还要对初始模糊图像进行去噪处理,而基于小波域

的软阈值和硬阈值去噪方法,其效果都不太理想,文[6]提出了一种基于小波系数的自适应去噪方法,可望推广本文方法的适用范围,还有待于进一步研究和实验。此外,去模糊后的图像的振铃效应虽然对视觉效果影响不大,但也需要降低或者消除,这也是需要改进的。

参考文献

- 1 王光学,曹长修,孙光宇,王健. 一种基于视觉特性的零树编码. 计算机工程与应用, 1999, 6
- 2 Kuo S-S, Johnston J D. Spatial noise shaping based on human visual sensitivity and its application to image coding. IEEE transaction on image processing, 2002, 11(5): 509~517
- 3 李树涛,王耀南,张昌凡. 基于视觉特性的多聚焦图像的融合. 电子学报, 2001, 12
- 4 肖忠,刘钊. 采用树特征平面的小波变换图像编码压缩方法. 通信学报, 1999, 20(11)
- 5 冉启文著. 小波变换与分数傅里叶变换理论及应用. 哈尔滨工业大学出版社, 2001, 3
- 6 Chang S G, Yu B, Vetterli M. Spatially adaptive wavelet thresholding with context modeling for image denoising. IEEE transaction on image processing, 2000, 9(9): 1522~1531