IBA 体系结构的研究*)

侯宗浩[1,2] 郑守淇[1] 董小社[1] 黄泳翔[1]

(西安交通大学新型计算机研究所 西安710049)[1] (第四军医大学京西医院信息科 西安710032)[2]

Research of Infiniband Architecture

HOU Zong-Hao ZHENG Shou-Qi DONG Xiao-She HUANG Yong-Xiang

(NeoComputing Research Institution, Xi'an JiaoTong University, Xi'an 710049)[1]

(Department of Information Technology, Xi-jing Hospital, Xi'an 710032)[2]

Abstract Infiniband Architecture is a new, network-centered, switch-based network technology. It is mainly designed for Internet Data Center. Abandoning the share bus technology, IBA carries out the redundant connections among network, storage and servers by a group of switches. By the mechanism provided by IBA, the data in the buffers of the two processes, which are in different nodes on the network, can be exchanged, reducing the overhead of the traditional network protocol, minimizing the switch of the context and promoting the performance of the system. This paper presents the current problem faced by cluster. We mainly stress the analyzing and the research of the topology and the mechanism of communication of IBA. At last, we talk about our next work on the research of this technology in this paper.

Keywords IBA, NAS, SAN, VIA

1 引言

塔式服务器中存在三个层次的总线:即系统内存总线,主机 I/O 总线和主机 I/O 控制器,系统内存总线提供内存和各 CPU 之间的通讯,并通过桥连向主机 I/O 总线;主机 I/O 总线通过主机 I/O 控制器(HBA)连向外部设备;HBA 是位于扩展卡上或集成于主板的一个集成芯片,如:SCSI、FC 以及以太网等。HBA 往往连接多个存储设备,因此各设备之间实际上是共享 I/O 总线的。

这种共享总线结构有两个明显的不足:其一,限定了所有的设备只能分享指定的带宽,从而连接的设备越多,争用总线的冲突就越多,每个设备所能得到的带宽就越小;虽然可采用多重并行信号提高总线的带宽,但是相应增加的针脚数量,使总线占据的空间也变大;若通过提升总线工作频率来提高带宽,就要缩短信号的有效距离。当这个距离缩短到以厘米计时,主板的设计将变得十分复杂,随着设备的增加,信号周期的判定和终结也会变得越来越困难,直接影响到了系统的性能和稳定性。

另一个不足就是其"外挂"能力,这种结构存储器隶属于特定服务器,外接存储器的最大数目在服务器设计时就已经被限定,因而限制了存储容量的扩展。

当前普遍使用的集群服务器就是通过一组基于塔式的应用服务器来分散负载,利用相对便宜的冗余系统来获得相对较高的计算性能和存储容量,然而,这种传统集群服务器,继承了塔式服务器的缺点:第一,当要共享某个数据时,首先必须竞争共享的总线,因此改变计算机系统对存储设备严密控制的结构,将存储设备从一个以主机为中心的结构转化为一个以网络为中心的新的结构体系成为未来的发展方向,如存储域网(storage area networks,SAN)和网络连接存储(network-attached storage,NAS)。

第二是使用的网络协议,目前,主流集群服务器大多数基于 TCP/IP 协议,此协议在设计中过多地考虑了网络的不可靠性,因此产生了许多对局域网来说不必要的开销。

IBA 主要针对 IDC(Internet Data Center)和局域网而设计,它基于这样的前提:网络线路可靠、延迟小、吞吐率高;本地服务器使用同一类操作系统和文件系统,共享有限数量的应用服务器,地理上位于同一个数据中心,由单独一组管理员负责配置和管理。因此可以对网络协议的设计进行简化。

以 IBA 技术构建服务器,存储和计算资源可独自扩展;应用服务器和文件服务器具有很好的容错性;异构的、基于标准的硬件,降低了成本,易于管理。

2 IBA 拓扑结构分析

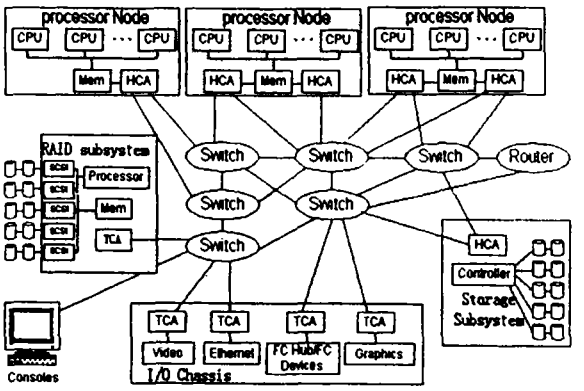

图1 IBA 拓扑结构

*)项目资助:863重点项目,新型网络服务器系统(01-2D01)。侯宗浩 博士生,研究方向为计算机体系结构。郑守淇 博士生导师,研究方向为计算机体系结构。董小社 副教授,研究方向网格计算。黄泳翔 硕士生,研究方向为并行文件系统。

如图1[8]，IBA 的核心是一个基于交换的通信子网 Fabric。Fabric 由主机通道适配器（Host Channel Adapter，HCA）、目标通道适配器（Target Channel Adapter，TCA）和可堆叠的 IBA 交换机（Switch）组成。HCA 靠近服务器的 CPU 和内存子系统，负责服务器与系统核心之间的连接和数据传输；TCA 靠近存储子系统和其它外部设备，负责 IBA 的数据流与传统的外部 I/O 控制器（SCSI、以太网等等）之间的通信；IBA 交换机位于 HCA 和 TCA 之间，负责在两者之间传递数据包。一个 HCA 可以同多个 TCA 建立连接。

对 IBA 的分析如下：

交换结构 IBA 以网络为核心，通过一组可堆叠的交换机在远程存贮器、网络以及服务器等设备之间建立点对点的连接链路，将所有的 I/O 设备和主机设备连接起来。因此 IBA 能够支持多种拓扑结构：点对点，多点对多点，从而减少 I/O 联机、增强驱动能力，缓解了各硬件设备之间的数据流量拥塞。

存储结构 IBA 使存储子网和应用服务器相分离，存储系统的管理和扩展独立于存储系统所连接的应用程序、操作系统和计算机的体系结构；IBA 打破了传统的存储设备和计算机之间的连接方式，放弃了计算机系统对存储设备进行严密控制的结构，存储设备成为网络的一个组成部分，而不再是某个特定系统的一部分，实现了真正意义上的存储共享。

瘦服务器 去掉了外存的服务器节点硬件更加紧凑，所用操作系统也比以前简洁，运算速度相对变快。

模块化 IBA 的每个组件都以模块方式实现，它定义了各模块几种统一的标准尺寸、机电性能和软硬件接口，这样保证了各厂家产品的交叉使用和产品的历史兼容性，使服务器结构从高密度机箱（如机架）向高度模块化的服务器阵列过渡。应用模块化，从而使主机系统规模变小，外围设备不再局限于系统内部，可以方便地进行系统配置。

冗余技术 IBA 的每个节点可连接一到多个交换机，任意两个节点之间可存在多个路径。当一个链路失效，其工作可自动切换到其它网络节点。可方便地实现服务器和存储系统的冗余，要增加更多的 TCA 节点或终端只需简单地添加并设定一下 IBA 的交换路由。这种设计提高了容错途径，也增加了传输的带宽。

串行总线结构 并行总线一般基于共享总线结构，串行标准一般基于网络结构，因此在高端系统上，串行总线结构的优势比较明显。

HCA 和 TCA 之间通过双向串行链路实现全双工通信，为此定义了四条线缆，形成一组发送和接收的线缆对，以实现数据并发的发送和接收。每组单向传输（一对线缆）速率可达 2.5Gb/s。通过增加双向串行链路组的数目，可以动态地提高链路的理论带宽。IBA 定义了 1x、4x、12x 三组链路标准，对应的线缆数目分别为4、16、48，因此，双向速率分别可达5Gb/s、20b/s、60Gb/s，随着 IB 技术的发展和信号传输率的增加，每个链路上的带宽将进一步增加。由于每个链路的有效带宽（链路层采用8B/10B 编码）大概可达到80%，由此可得出具体有效吞吐率。描述如下：

链路宽度	单向速率	可用带宽	有效传输的吞吐率
1x	2.5Gb/s	2Gb/s(250MB/s)	(250+250)MB/s
4x	10Gb/s	8Gb/s(1MB/s)	(1+1)GB/s
12x	30Gb/s	24Gb/s(3MB/s)	(3+3)MB/s

远程直接内存访问 IBA 采用了内存-内存网络互连技术，在数据传输过程中，可以旁路主机操作系统，直接实现两个通信进程用户缓存之间内容的传输。

精心设计的管理机制 IBA 定义了统一的机群管理设施来实现对本地子网的配置并确保其能连续运行。通过分布在端口上的子网管理代理（Subnet Management Agent，SMA），子网管理器（Subnet Manager，SA）可以完成拓扑发现、子网配置（配置交换机中的路由表、设置交换机和通道适配器中的操作参数如：子网 ID，M_Keys，P_Keys，SL 到 VL 的映射等）和子网维护功能；通过散布在节点和断口上的通用服务代理（General Services Agents，GSAs），通用服务管理器可以执行各种与性能、通讯和 I/O 设备相关的管理功能。通用服务管理包括以下几个：子网管理（Subnet Administrator）、性能管理（Performance Management）、基板管理（Baseboard Management）、通信管理（Communications Management）、设备管理（Device Management）、SNMP 隧道（SNMP Tunneling）、供应商专用管理（Vendor-specific Management）、专用应用的管理（Application-specific Management）等。

3 IBA 内部通信机制分析

3.1 IBA 层次体系

与传统 TCP/IP 网络相同，IBA 也分为不同的层，每一层的协议独立于其他层，下层为相邻的上层提供服务。如图2[1]所示。

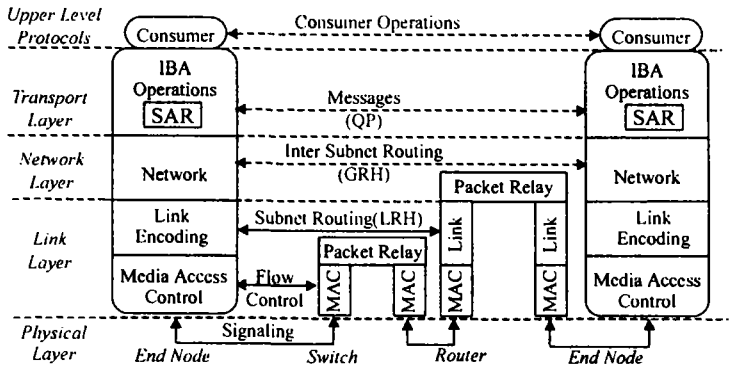

图2 IBA 层次体系

每一层定义了各自独立的数据包域（图3[1]）：

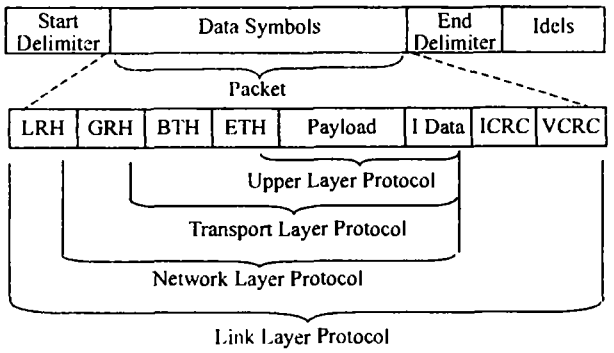

图3 IBA 数据包结构

包通信方面，IBA 架构分为数据传输和网络管理两种包类型：网络管理数据包主要用于控制设备计数、子网定向和容错等；子网管理器以16位的本地标识地址(LID，见下)对这两种包的路径和交换属性进行管理。

物理层负责实际的数据传输。包括定义三种链路速率，铜线和光纤等物理介质，以及用于光纤和铜线的标准连接器和电缆，用于机架系统的背板连接器。

链路层提供基本的子网连接，维护着数据在子网内的高效传输。服务质量(Quality of Service，QoS)主要由该层体现，即虚拟路径(Virtual Lane，VL)的使用。子网内的数据包都通过链路层进行传送，链路层对每个数据包实现两种循环冗余检测(CRC)，即可变 CRC(Variant CRC，VCRC)和固定 CRC (Invariant CRC ，ICRC)。VCRC 提供了两次转发间链路级的数据完整性，ICRC 提供了端到端的数据完整性。

网络层提供了子网间数据包的路由。每个经过路由器的数据包都被分配一个全局路由头(Global Route Header，GRH)和一个128位的 IPv6地址。每个子网都有一个64位的统一的子网前缀，子网间通过这一前缀进行路由。

传输层包括很多关键技术，其中包括包的按顺序发送、分割、信道的复用和基本传输服务。一些基本的网络参数，如最大传输单元(Maximum Transfer Unit，MTU)的大小也由这个隐藏的传输层处理。为了简单起见，这种传输的绝大部分特性都和目前的网络技术类似。

链路层和传输层定义、交换和传输 InfiniBand 数据包，并且要保证负载平衡及高性能的服务。网络层负责网间的数据交换，物理层提供了硬件组件。这种设计方式，可以把操作系统和底层的硬件分开，用户无需直接和硬件通信，而是通过几个中间层。

3.2 包的管理与通讯机制

IBA 借鉴了 VIA 的通讯机制。VIA 是由 Intel、Microsoft 和其他技术公司共同开发的用户层的互连协议标准，目的是避免传统网络协议的过度开支和延迟。VIA 为每个应用进程提供了受保护的、直接访问网络硬件的虚拟接口(Virtual Interface，VI)。

VI 用户访问网络硬件资源时，由 VI 用户代理(VI User Agent)先向内核登记用户缓存，随后将控制转给 VI 核心代理。VI 核心代理将应用程序所划出的缓存的控制句柄交给 VI 网络适配器。这些缓存通过队列来分配，而队列由驱动程序和 VI 网络适配器来管理，支持 VI 的应用程序使用这些队列在系统之间读写数据。VI 应用程序要发送数据，它首先将在数据存储位置的指针等信息组成一个描述符送进 VI 的队列对中，VI 适配器在后台对其进行读、写等 RDMA(一个系统去更新另一个系统的内存)操作。这样就避免了传统协议在数据通过网络协议栈时所做的多次数据拷贝和上下文切换的开销，节约了时间，实现了零拷贝。如图4[3]所示：

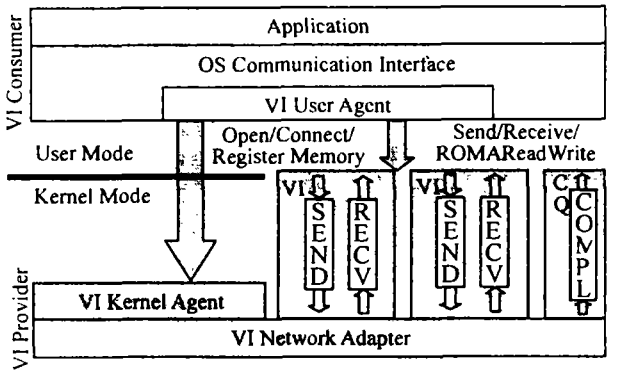

图4 VI 体系结构

3.3 通讯接口队列对(Queue Pairs，QPs)

HCA 和 TCA 定义了逻辑接口 QP。链路的每一端都包含一些要发往另一端的消息队列，根据应用的不同，可对每一个 QP 设置不同的服务级别(Service Level，SL)。如多媒体视频数据需要提供连续的具有事件同步消息的数据流。

QP 由发送工作队列、接收工作队列组成。发送队列保存着把本地数据发往另一个远程进程的指令，接收队列保存着把远程数据放到本地某个地址的指令。

QP 屏蔽了硬件实现细节，它们各自独立地进行操作。系统为用户分配一个 QP 并指定其服务级别，即提供了一个虚拟的通讯端口。建立连接主要完成：双方队列对的绑定，设置目标地址、服务级别等内容，并协商一些限制。IBA 中的所有节点都提供 QP0 和 QP1，用于管理接口。

一个通道适配器可以提供多达2^{24}个 QP。适配器给每个 QP 分配一个队列对号(QPN)，作为其唯一的标识符。数据包中包含目标 QP 的 QPN，当适配器收到一个包时，它根据包中目的 QPN 对包进行处理。

3.4 数据传输单元(Data transfer units)

IBA 既可用数据报也可用面向连接的方式进行通讯，因此 IBA 结构既可以使用块传输设备，也可以使用连续的数据流设备。通信的基本单位是消息，硬件自动提供消息的分段和重组，发送的消息被分成包，在另一端，多个包又组成一个消息。数据传输层提供六种通讯协议：可靠性连接、不可靠性连接、可靠性数据报、不可靠性数据报、多播连接和源数据包(raw packets)，前四个协议是必需的，多播连接和源数据包在设备开发时可视具体情况决定是否支持。

在 IBA 中，可靠性意味着硬件对每一个包维护一个顺序号，每一个接收的包都向数据源发一个确认消息，同时拒绝重复包、向丢包的数据源要求重发，从而对 fabric 失效引起的问题提供简单的恢复。

3.5 虚链路(Virtual Lanes，VLs)

为了提供流控，HCA 和 TCA 之间的连接被划分为一系列 VLs，每一个 VL 内部，包的流动具有不同的优先级，每个端口(port)的 VL 代表了一组发送和接收缓冲区，每个端口可支持16个 VL，其中：VL0到 VL14为数据虚链路，每个端口至少支持 VL0，VL15用于子网管理，所有端口都必须支持。

端口(Port)对流过每个虚链路的数据独立地进行控制，以保证一个虚链路的过载不会影响到其它虚链路，通过使用

不同的虚链路，可以将不同的设备通讯组分开以防止相互干扰。如，一个300 MBs 的链路被分成15个虚链路，每个虚链路可分配20M 的带宽，可以为15对设备提供服务而互不干扰。

VL 的指定只是端口之间的事情，不需穿过整个 fabric，包有16个服务等级(Service Level，SL)，由包头进行设置，当包在 fabric 中传输时，根据服务级别来确定具体使用哪一个 VL，每一个端口维护着一个由子网管理器设置的 SL 到 VL 的映射表，用来确定要发送的包将送往哪一个 VL。当一个链路两端的端口 的 VL 数目不同时，数目多的端口将会减少 VL 数目以便互相匹配，对仅支持一个数据虚链路的端口来说，所有的数据流向 VL0。

3.6 分区(partitions)

Fabric 可以为特定的分区保留某些 SLs，从而子网管理器(SM)从这些分区之间分离出业务流，即使两个分区具有相同的 QoS ，也能保证每一个分区公平地分享自己的带宽而不受其他分区的影响。

3.7 密钥(keys)

IBA 使用各种密钥提供隔离和保护功能，密钥由 administrative 实体指定，以各种方式用于消息之中。密钥包括：管理密钥(Management Key ，M_Key)、基板管理密钥(Baseboard Management Key，B_Key)、分区密钥(Partition Key，P_Key)、队列密钥(Queue Key，Q_Key)、内存密钥(Memory Keys，L_Key and R_Key)。

3.8 编址 (Addressing)

每个端节点可包含一到多个适配器，每个适配器可包含一到多个端口，每个端口包含一个 LID 和 GID，IBA 中的节点还包括一个 GUID 用于管理。下面通过对各个地址功能的分析来了解这种管理：

LIDs 是子网管理器给每个端口分配的一个局部的16位的标识符，在子网内部，LIDs 是唯一的，子网内的交换机使用它来路由数据包。交换机中的路由表由子网管理器根据节点端口的 LIDs 和它与交换机的相对位置来设置，每个数据包包含有源 LID (SLID)和目的 LID (DLID)，分别表示产生和接受数据包的端口号。

LMC 指定了 LID 的不重要的位的数目，TCA 的物理端口在验证数据包的 DLID 时忽略这几位的差别，而交换机将对这些位加以区分，这样，通道适配器中的一个端口在 fabric 中可变成2^{LMC}个端口。从而子网管理器可在 fabric 中对同一对连结安排不同的路径。例如：如果子网中两个端口 A、B 之间存在4条路径，B 端口的 LID 为4，LMC 为2，则交换机将把它看成4个 LIDs：{4，5，6，7}，同样，如果 A 端口的 LID 为 8，LMC 为2，则 LIDs 为 {8，9，10，11}。

GID 每个端口最少有一个全局标识符 GID (Global Identifier)。GID 是一个全球唯一的 IPv6地址。每一个网间传输的数据包必须包含 GRH(Global Route Header)(GRH 指定了源 GID(SGID)和目标 GID (DGID)，分别代表了网间传输的数据包的源和目的地址)。GID 是当路由器在各子网之间转发包时使用，与交换机没有关系。

GUID (子网前缀＋端口 GUID 形成端口的 GID)

GUID(Globally Unique Identifier)实际上就是一个 EUI-64地址，它是由 IEEE 引入的一种新类型的64位的 MAC 地址，它与 MAC 地址的区别在于，在旧的 MAC 地址中，厂家的选择位为24，而 EUI-64则为40位，通过相应的处理，兼容原有的 MAC 地址。

为了便于管理，每一个端节点和交换机有一个 GUID；适配器的开发商为每一个连向端节点的适配器分配一个 GUID，称为节点 GUID，它不同于 GID；一台服务器可能包含多个 HCAs，以提供冗余或连向另一个 fabrics，每个 HCA 拥有自己的 GID；适配器的销售商为适配器上的每一个端口指定一个端口 GUID，端口 GUID 和子网 ID(也叫子网前缀)组合在一起形成端口的 GID。

综上所述，GID 是一个全球唯一的128位 IPv6 地址，LID 是仅用于子网内部路由的逻辑地址，子网管理器提供节点 GUID 到 LID/GID 的名字的服务。QP 的地址由 GID ＋ LID ＋ QPN 组成。

4 相关工作

以网络为中心的存储使服务器结构发生了很大的变化，其影响是深远的，需要研究的内容也很多，包括：设计支持 IBA 的网络硬件如：服务节点、交换机、HCA、TCA、路由器、接口机以及这些硬件所需要的驱动程序；改造现有操作系统、应用程序和数据库以支持 IBA；设计与实现基于 IBA 的集群操作系统；开发子网管理软件、改造相关外设以支持或包含 TCA。

另外，和普通机群一样，解决 RASUM 问题是 IBA 机群的目标之一。RASUM 即：可靠性(RELIABILITY)、有效性(AVAILABILITY)、可服务性(SERVICEABILITY)、可用性(USABILITY)、可管理性(MANAGEABILITY)。

结论 IBA 作为一种新型网络体系结构获得了业界的广泛支持。但作为一种全新的接口规范，IBA 定义了自己的硬件标准，和目前的软、硬件标准不兼容。它的软件协议需要操作系统、驱动程序和硬件接口的全面支持，也会涉及服务器应用程序。另外 IBA 体系中引入了 SAN、NAS 和 RDMA 技术，提出了许多有待解答的问题，因此早期的 IBA 系统将通过软件中间层来实现以保持与传统软硬件和应用的兼容，当然，随着技术的成熟，这些工作将逐渐全部由硬件来完成。

作为一种全新的架构，IBA 成熟完善所需要的时间很可能比厂商估计的要长一些。IBA 技术的改进也将持续多年。随着 PCI 总线的局限性进一步被大家所认识，这种体系将逐渐成为市场的主流。

参 考 文 献

1 InfiniBand Architecture，Volume 1 - General Specifications，InfiniBand Trade Association，2000

2 InfiniBand Architecture，Volume 2 - Mechanical，Cable and Optical Specifications，InfiniBand Trade Association，2000

3 Compaq Computer Corp.，Intel Corporation，Microsoft Corporation. Virtual Interface Architecture Specification，Version 1.0. Dec. 1997

4 Baker M. University of Portsmouth，UK，Cluster Computing White Paper，Status-Final Release，Version 2.0，Dec. 2000

5 Futral W T. Infiniband Architecture Development and Deployment -a strategic Guide to Server I/O Solutions，Intel Press，2001

6 InfiniBand Trade Association. http：//www.infinibandta.org

7 Virtual Interface Architecture organization. http：//www.viarch.org

8 InfiniBand Architecture：Next-Generation Server I/O ，VECTORS WHITE PAPER，DELL，Oct. 2000 http：//www.dell.com/us/en/arm/topics/vectors_2000-infiniband.htm