



湖南大学
HUNAN UNIVERSITY

《 计算机系统设计 》

题 目	<u>Infiniband 网络结构分析</u>
专 业	<u>智能科学与技术</u>
班 级	<u>智能 1702</u>
姓 名	<u>唐昌均</u>
学 号	<u>201708010712</u>

Infiniband 开放标准技术简化并加速了服务器之间的连接,同时支持服务器与远程存储和网络设备的连接。

IB 技术的发展

1999 年开始起草规格及标准规范, 2000 年正式发表, 但发展速度不及 Rapid I/O、PCI-X、PCI-E 和 FC, 加上 Ethernet 从 1Gbps 进展至 10Gbps。所以直到 2005 年之后, InfiniBand Architecture(IBA)才在集群式超级计算机上广泛应用。全球 Top 500 大效能的超级计算机中有相当多套系统都使用上 IBA。随着越来越多的大厂商正在加入或者重返到它的阵营中来, 包括 Cisco、IBM、HP、Sun、NEC、Intel、LSI 等。InfiniBand 已经成为目前主流的高性能计算机互连技术之一。为了满足 HPC、企业数据中心和云计算环境中的高 I/O 吞吐需求, 新一代高速率 56Gbps 的 FDR (Fourteen Data Rate) 和 EDR InfiniBand 技术已经出现。

IB 技术的优势

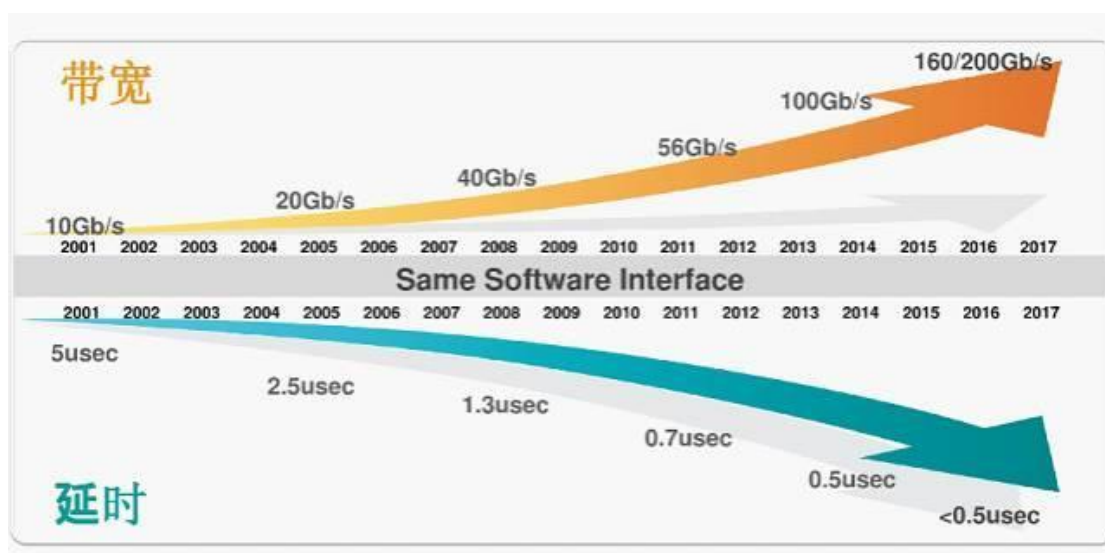
Infiniband 大量用于 FC/IP SAN、NAS 和服务器之间的连接,作为 iSCSI RDMA 的存储协议 iSER 已被 IETF 标准化。目前 EMC 全系产品已经切换到 Infiniband 组网, IBM/TMS 的 FlashSystem 系列, IBM 的存储系统 XIV Gen3, DDN 的 SFA 系列都采用 Infiniband 网络。

相比 FC 的优势主要体现在性能是 FC 的 3.5 倍, Infiniband 交换机的延迟是 FC 交换机的 1/10, 支持 SAN 和 NAS。

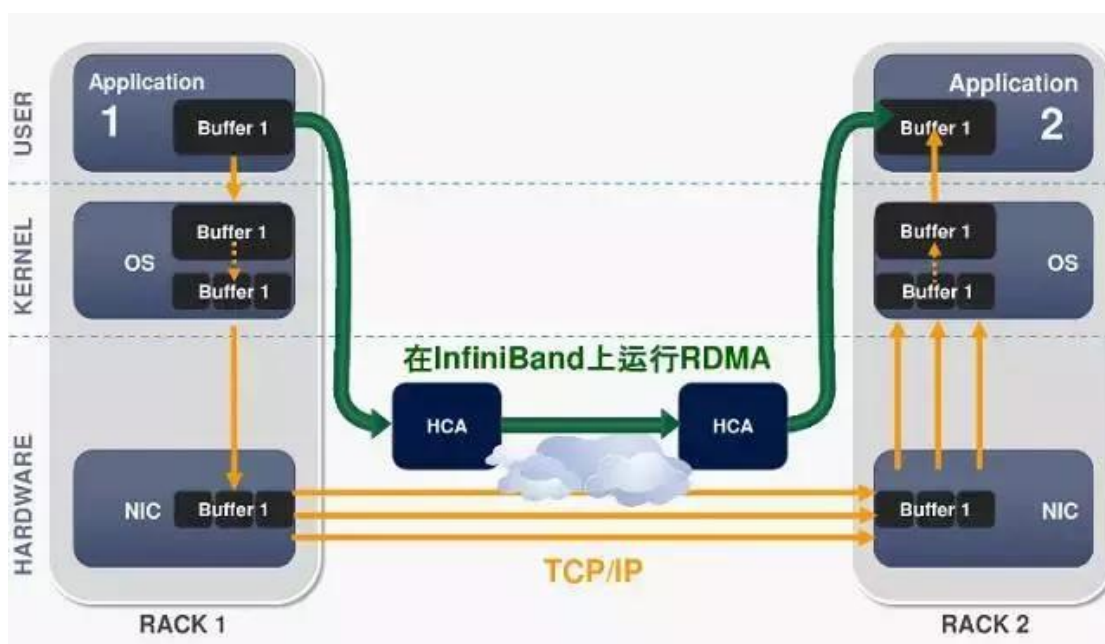
存储系统已不能满足于传统的 FC SAN 所提供的服务器与裸存储的网络连接架构。HP SFS 和 IBM GPFS 是在 Infiniband fabric 连接起来的服务器和 iSER Infiniband 存储构建的并行文件系统, 完全突破系统的性能瓶颈。

Infiniband 采用 PCI 串行高速带宽链接, 从 SDR、DDR、QDR、FDR 到 EDR HCA 连接, 可以做到 1 微妙、甚至纳米级别极低的时延, 基于链路层的流控机制实现先进的拥塞控制。

InfiniBand 采用虚通道(VL 即 Virtual Lanes)方式来实现 QoS, 虚通道是一些共享一条物理链接的相互分立的逻辑通信链路, 每条物理链接可支持多达 15 条的标准虚通道和一条管理通道(VL15)。



RDMA 技术实现内核旁路，可以提供远程节点间 RDMA 读写访问，完全卸载 CPU 工作负载，基于硬件传出协议实现可靠传输和更高性能。



相比 TCP/IP 网络协议，IB 使用基于信任的、流控制的机制来确保连接的完整性，数据包极少丢失，接受方在数据传输完毕之后，返回信号来标示缓存空间的可用性，所以 IB 协议消除了由于原数据包丢失而带来的重发延迟，从而提升了效率和整体性能。

TCP/IP 具有转发损失的数据包的能力，但是由于要不断地确认与重发，基于这些协议的通信也会因此变慢，极大地影响了性能。

IB 基本概念

IB 是以通道为基础的双向、串行式传输，在连接拓扑中是采用交换、切换式结构(Switched Fabric)，在线路不够长时可用 IBA 中继器(Repeater)进行延伸。每一个 IBA 网络称为子网(Subnet)，每个子网内最高可有 65,536 个节点(Node)，IBA Switch、IBA Repeater 仅适用于

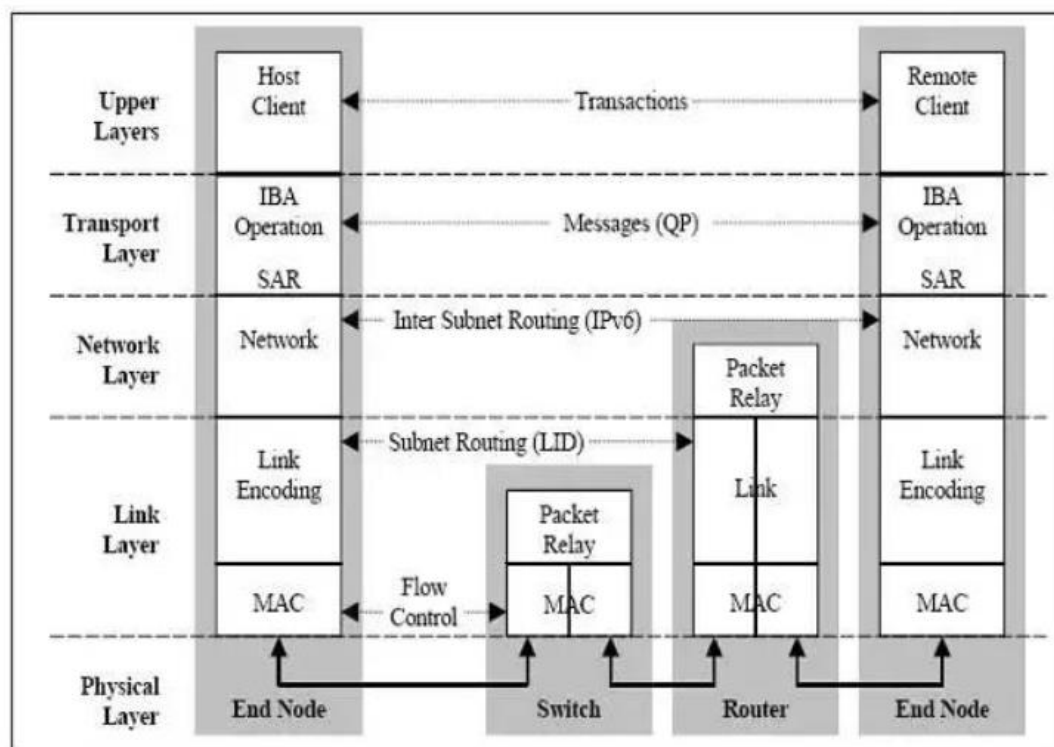
Subnet 范畴，若要通跨多个 IBASubnet 就需要用到 IBA 路由器(Router)或 IBA 网关器(Gateway)。

每个节点(Node) 必须透过配接器(Adapter)与 IBA Subnet 连接，节点 CPU、内存要透过 HCA(Host Channel Adapter)连接到子网；节点硬盘、I/O 则要透过 TCA(Target Channel Adapter)连接到子网，这样的一个拓扑结构就构成了一个完整的 IBA。

IB 的传输方式和介质相当灵活，在设备机内可用印刷电路板的铜质线箔传递(Backplane 背板)，在机外可用铜质缆线或支持更远光纤介质。若用铜箔、铜缆最远可至 17m，而光纤则可至 10km，同时 IBA 也支持热插拔，及具有自动侦测、自我调适的 Active Cable 活化智能性连接机制。

IB 协议简介

InfiniBand 也是一种分层协议(类似 TCP/IP 协议)，每层负责不同的功能，下层为上层服务，不同层次相互独立。IB 采用 IPv6 的报头格式。其数据包报头包括本地路由标识符 LRH，全局路由标示符 GRH，基本传输标识符 BTH 等。



1、物理层

物理层定义了电气特性和机械特性，包括光纤和铜媒介的电缆和插座、底板连接器、热交换特性等。定义了背板、电缆、光缆三种物理端口。

并定义了用于形成帧的符号(包的开始和结束)、数据符号(DataSymbols)、和数据包直接的填充(Idles)。详细说明了构建有效包的信令协议，如码元编码、成帧标志排列、开始和结束定界符间的无效或非数据符号、非奇偶性错误、同步方法等。

2、链路层

链路层描述了数据包的格式和数据包操作的协议，如流量控制和子网内数据包的路由。链路层有链路管理数据包和数据包两种类型的数据包。

3、网络层

网络层是子网间转发数据包的协议，类似于 IP 网络中的网络层。实现子网间的数据路由，数据在子网内传输时不需网络层的参与。

数据包中包含全局路由头 GRH，用于子网间数据包路由转发。全局路由头部指明了使用 IPv6 地址格式的全局标识符(GID)的源端口和目的端口，路由器基于 GRH 进行数据包转发。GRH 采用 IPv6 报头格式。GID 由每个子网唯一的子网标识符和端口 GUID 捆绑而成。

4、传输层

传输层负责报文的分发、通道多路复用、基本传输服务和处理报文分段的发送、接收和重组。传输层的功能是将数据包传送到各个指定的队列(QP)中，并指示队列如何处理该数据包。当消息的数据路径负载大于路径的最大传输单元(MTU)时，传输层负责将消息分割成多个数据包。

接收端的队列负责将数据重组到指定的数据缓冲区中。除了原始数据报外，所有的数据包都包含 BTH，BTH 指定目的队列并指明操作类型、数据包序列号和分区信息。

5、上层协议

InfiniBand 为不同类型的用户提供了不同的上层协议，并为某些管理功能定义了消息和协议。InfiniBand 主要支持 SDP、SRP、iSER、RDS、IPoIB 和 uDAPL 等上层协议。

SDP(SocketDirect Protocol)是 InfiniBand Trade Association (IBTA)制定的基于 infiniband 的一种协议，它允许用户已有的使用 TCP/IP 协议的程序运行在高速的 infiniband 之上。

SRP(SCSIRDMA Protocol)是 InfiniBand 中的一种通信协议，在 InfiniBand 中将 SCSI 命令进行打包，允许 SCSI 命令通过 RDMA(远程直接内存访问)在不同的系统之间进行通信，实现存储设备共享和 RDMA 通信服务。

iSER(iSCSIRDMA Protocol)类似于 SRP(SCSI RDMA protocol)协议，是 IB SAN 的一种协议，其主要作用是把 iSCSI 协议的命令和数据通过 RDMA 的方式跑到例如 Infiniband 这种网络上，作为 iSCSI RDMA 的存储协议 iSER 已被 IETF 所标准化。

RDS(ReliableDatagram Sockets)协议与 UDP 类似，设计用于在 Infiniband 上使用套接字来发送和接收数据。实际是由 Oracle 公司研发的运行在 infiniband 之上，直接基于 IPC 的协议。

IPoIB(IP-over-IB)是为了实现 INFINIBAND 网络与 TCP/IP 网络兼容而制定的协议，基于 TCP/IP 协议，对于用户应用程序是透明的，并且可以提供更大的带宽，也就是原先使用 TCP/IP 协议栈的应用不需要任何修改就能使用 IPoIB。

uDAPL(UserDirect Access Programming Library)用户直接访问编程库是标准的 API，通过远程直接内存访问 RDMA 功能的互连（如 InfiniBand）来提高数据中心应用程序数据消息传送性能、伸缩性和可靠性。

IB 应用场景

Infiniband 灵活支持直连及交换机多种组网方式，主要用于 HPC 高性能计算场景，大型数据中心高性能存储等场景，HPC 应用的共同诉求是低时延(<10 微秒)、低 CPU 占有率(<10%) 和高带宽(主流 56 或 100Gbps)。一方面 Infiniband 在主机侧采用 RDMA 技术释放 CPU 负载，可以把主机内数据处理的时延从几十微秒降低到 1 微秒；另一方面 InfiniBand 网络的高带宽(40G、56G 和 100G)、低时延(几百纳秒)和无丢包特性吸取了 FC 网络的可靠性和以太网的灵活扩展能力。