**Infiniband网络结构分析**

**目录：**

**一、Infiniband简介**

**二**、**InfiniBand架构**

三、**InfiniBand基本组件**

**四、上层协议**

五、**InfiniBand技术的优势**

**一、Infiniband简介**

InfiniBand（直译为“无限带宽”技术，缩写为IB）是一个用于[高性能计算](https://baike.baidu.com/item/%E9%AB%98%E6%80%A7%E8%83%BD%E8%AE%A1%E7%AE%97" \t "_blank)的计算机网络通信标准，它具有极高的[吞吐量](https://baike.baidu.com/item/%E5%90%9E%E5%90%90%E9%87%8F)和极低的[延迟](https://baike.baidu.com/item/%E5%BB%B6%E8%BF%9F)，用于计算机与计算机之间的数据互连InfiniBand也用作服务器与存储系统之间的直接或交换互连，以及存储系统之间的互连。

与目前计算机的I/O子系统不同，InfiniBand是一个功能完善的网络通信系统。InfiniBand贸易组织把这种新的总线结构称为I/O网络，并把它比作开关，因为所给信息寻求其目的地址的路径是由控制校正信息决定的。InfiniBand使用的是网际协议版本6的128位地址空间，因此它能提供近乎无限量的设备扩展性。

通过InfiniBand传送数据时，数据是以数据包方式传输，这些数据包会组合成一条条信息。这些信息的操作方式可能是远程直接内存存取的读写程序，或者是通过信道接受发送的信息，或者是多点传送传输。就像大型机用户所熟悉的信道传输模式，所有的数据传输都是通过信道适配器来开始和结束的。每个处理器（例如个人电脑或数据中心服务器）都有一个主机通道适配器，而每个周边设备都有一个目标通道适配器。通过这些适配器交流信息可以确保在一定服务品质等级下信息能够得到有效可靠的传送。

**二**、**InfiniBand架构**

InfiniBand采用双队列程序提取技术，使应用程序直接将数据从适配器送入到应用内存（称为远程直接存储器存取或RDMA）, 反之依然。在TCP/IP协议中,来自网卡的数据先拷贝到核心内存，然后再拷贝到应用存储空间，或从应用空间将数据拷贝到核心内存,再经由网卡发送到Internet。这种I/O操作方式，始终需要经过核心内存的转换，它不仅增加了数据流传输路径的长度，而且大大降低了I/O的访问速度，增加了CPU的负担。而SDP则是将来自网卡的数据直接拷贝到用户的应用空间，从而避免了核心内存参与。这种方式就称为零拷贝，它可以在进行大量数据处理时，达到该协议所能达到的最大的吞吐量。

InfiniBand的协议采用分层结构，各个层次之间相互独立，下层为上层提供服务。其中物理层定义了在线路上如何将比特信号组成符号，然后再组成帧、数据符号以及包之间的数据填充等，详细说明了构建有效包的信令协议等；链路层定义了数据包的格式以及数据包操作的协议，如流控、 路由选择、编码、解码等；网络层通过在数据包上添加一个40字节的全局的路由报头（Global Route Header, GRH）来进行路由的选择，对数据进行转发。在转发的过程中，路由器仅仅进行可变的CRC校验,这样就保证了端到端的数据传输的完整性；传输层再将数据包传送到某个指定的队列偶（Queue Pair, QP）中，并指示QP如何处理该数据包以及当信息的数据净核部分大于通道的最大传输单元MTU时，对数据进行分段和重组。

三、**InfiniBand基本组件**

InfiniBand的网络拓扑结构，其组成单元主要分为四类

1.HCA（Host Channel Adapter），它是连接内存控制器和TCA的桥梁

2.TCA(Target Channel Adapter)，它将I/O设备（例如网卡、SCSI控制器）的数字信号打包发送给HCA

3. InfiniBand link，它是连接HCA和TCA的光纤，InfiniBand架构允许硬件厂家以1条、4条、12条光纤3种方式连结TCA和HCA

4.交换机和路由器

无论是HCA还是TCA，其实质都是一个主机适配器，它是一个具备一定保护功能的可编程DMA（Direct Memory Access，直接内存存取 ）引擎。

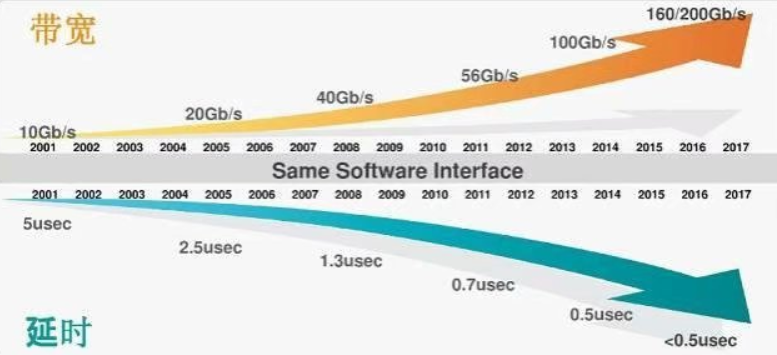
**四、上层协议**

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。

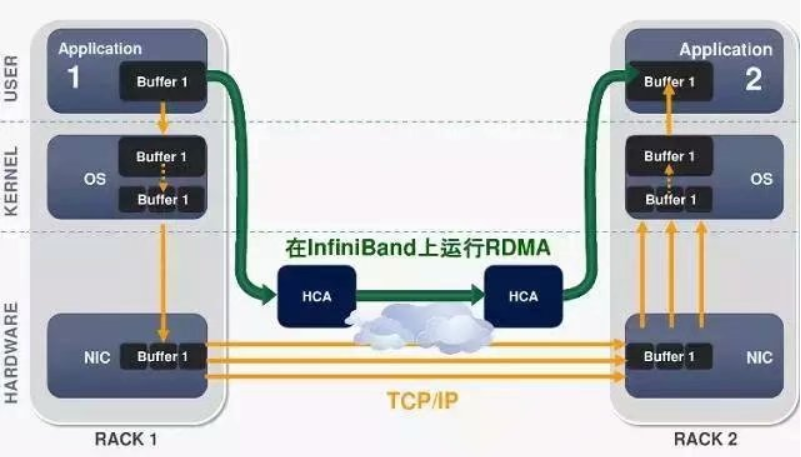
* SDP(SocketsDirect Protocol)是InfiniBand Trade Association (IBTA)制定的基于infiniband的一种协议，它允许用户已有的使用TCP/IP协议的程序运行在高速的infiniband之上。
* SRP(SCSIRDMA Protocol)是InfiniBand中的一种通信协议，在InfiniBand中将SCSI命令进行打包，允许SCSI命令通过RDMA(远程直接内存访问)在不同的系统之间进行通信，实现存储设备共享和RDMA通信服务。
* iSER(iSCSIRDMA Protocol)类似于SRP(SCSI RDMA protocol)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。

五、**InfiniBand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。

六、**InfiniBand应用**

在高并发和高性能计算应用场景中，当客户对带宽和时延都有较高的要求时，可以采用IB组网：前端和后端网络均采用IB组网，或前端网络采用10Gb以太网，后端网络采用IB。由于IB具有高带宽、低延时、高可靠以及满足集群无限扩展能力的特点，并采用RDMA技术和专用协议卸载引擎，所以能为存储客户提供足够的带宽和更低的响应时延。

IB目前可以实现以及未来规划的更高带宽工作模式有（以4X模式为例）：

·SRD (Single Data Rate)：单倍数据率，即8Gb/s

·DDR (Double Data Rate)：双倍数据率，即16Gb/s

·QDR (Quad Data Rate)：四倍数据率，即32Gb/s

·FDR (Fourteen Data Rate)：十四倍数据率，56Gb/s

·EDR (Enhanced Data Rate)：100 Gb/s

·HDR (High Data Rate)：200 Gb/s

·NDR (Next Data Rate)：1000 Gb/s+