# Infiniband网络结构分析

一．Infiniband基本概念

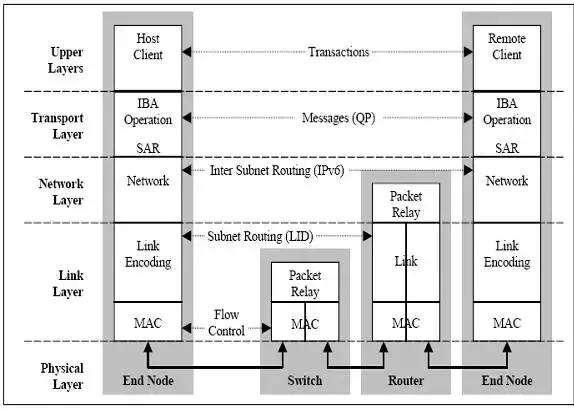
IB是以通道为基础的双向、串行式传输，在连接拓朴中是采用交换、切换式结构，在线路不够长时可用IBA中继器进行延伸。每一个IBA网络称为子网，每个子网内最高可有65,536个节点，IBA Switch、IBARepeater仅适用于Subnet范畴，若要通跨多个IBASubnet就需要用到IBA路由器或IBA网关器。

每个节点必须透过配接器与IBA Subnet连接，节点CPU、内存要透过HCA连接到子网；节点硬盘、I/O则要透过TCA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，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。

二．Infiniband协议简介

InfiniBand也是一种分层协议，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、和数据包直接的填充。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 标示符和端口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是InfiniBand Trade Association 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类似于SRP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。

三．Infiniband与其他协议的关系：   
　　作为总线的网络化，Infiniband有责任将其他进入服务器的协议在Infiniband的层面上整合并送入服务器。基于这个目的，今天Volatire已经开发了IP到Infiniband的路由器以及Fibre Channel到Infiniband的路由器。这样一来客观上就使得目前几乎所有的网络协议都可以通过Infiniband网络整合到服务器中去。这包括Fibre Channel, IP/GbE, NAS, iSCSI等等。另外2007年下半年Voltaire将推出万兆以太网到Infiniband的路由器。这里有一个插曲：万兆以太网在其开发过程中考虑过多种线缆形式。最后发现只有Infiniband的线缆和光纤可以满足其要求。最后万兆以太网开发阵营直接采用了Infiniband线缆作为其物理连接层。

四．Infiniband在存储中的地位：   
　　今天的Infiniband可以简单地整合Fibre Channel SAN、NAS以及iSCSI进入服务器。事实上除了作为网络化总线把其他存储协议整合进服务器之外，Infiniband可以发挥更大的作用。存储是内存的延伸，具有RDMA功能的Infiniband应该成为存储的主流协议。比较一下Infiniband和Fibre Channel我们可以看到Infiniband的性能是Fibre Channel的5倍，Infiniband交换机的延迟是Fibre Channel交换机的1/10。另外在构筑连接所有服务器和存储器的高速网络时使用Infiniband Fabric可以省去Fiber Channel Fabric,从而给客户带来巨大的成本节省。   
　　今天在使用Infiniband作为存储协议方面已经有了很大的进展。作为iSCSI RDMA的存储协议iSER已被IETF标准化。   
　　不同于Fibre Channel，Infiniband在存储领域中可以直接支持SAN和NAS。存储系统已不能满足于传统的Fibre Channel SAN所提供的服务器与裸存储的网络连接架构。Fibre Channel SAN加千兆以太网加NFS的架构已经严重限制了系统的性能。在这种情况下应运而生的则是由在Infiniband fabric连接起来的服务器和iSER Infiniband存储的基础架构之上的并行文件系统（诸如HP的SFS、IBM的GPFS等等）。在未来的服务器、存储器网络的典型结构将会是由Infiniband将服务器和Infiniband存储器直接连接起来，所有的IP数据网络将会通过万兆以太网到Infiniband的路由器直接进入Infiniband Fabric。   
　　在存储厂商方面Sun, SGI, LIS LOGIC,飞康软件等公司都已推出自己的Infiniband存储产品。在中国新禾科技公司也推出了他们的Infiniband存储系统。   
　　从价格的角度，今天的Infiniband是万兆以太网的几分之一。Inifiniabnd有比FibreChannel高5倍的性能，在价格上则已与Fibre Channel在同一个数量级上。

五．Infiniband**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。

