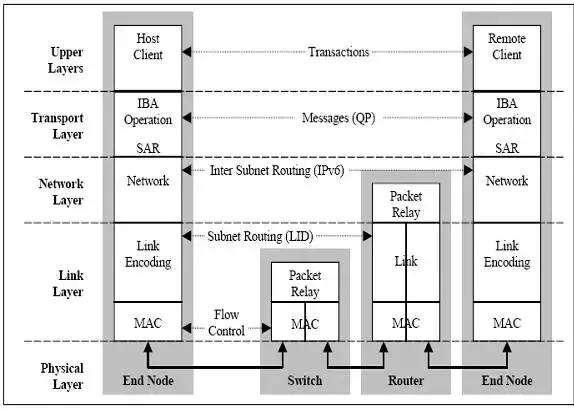
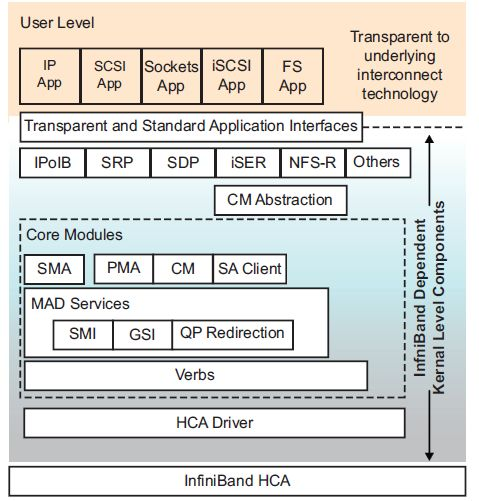
# InfiniBand分析

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(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。

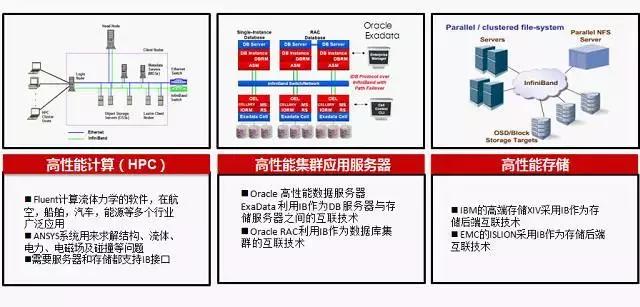
RDS(ReliableDatagram Sockets)协议与UDP 类似，设计用于在Infiniband 上使用套接字来发送和接收数据。实际是由Oracle公司研发的运行在infiniband之上，直接基于IPC的协议。

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。

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。