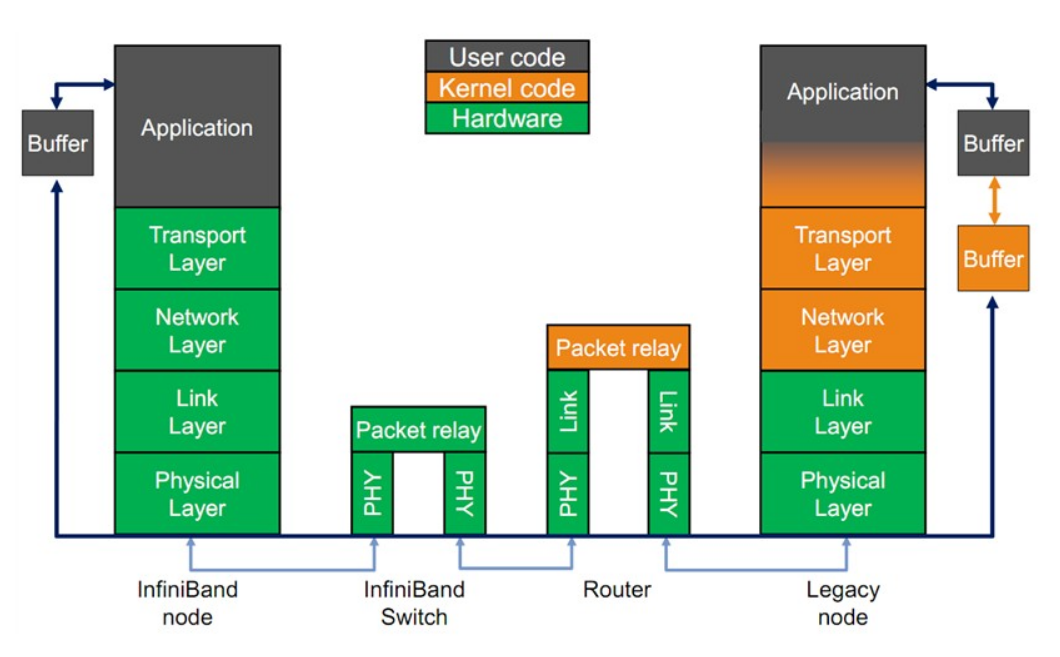
# infiniband网络架构分析

InfiniBand架构是一种支持多并发链接的“转换线缆”技术，它是新一代服务器硬件平台的I/O标准。由于它具有高带宽、低延时、 高可扩展性的特点，它非常适用于服务器与服务器（比如复制，分布式工作等），服务器和存储设备（比如SAN和直接存储附件）以及服务器和网络之间（比如LAN， WANs和the Internet）的通信 。

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)。

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(QueuePair,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。



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。

Infiniband的网络拓扑结构如图2，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

