# Infiniband网络结构分析

1. **Infiniband网络介绍**

InfiniBand是一种网络通信协议，它提供了一种基于交换的架构，由处理器节点之间、处理器节点和输入/输出节点(如磁盘或存储)之间的点对点双向串行链路构成。每个链路都有一个连接到链路两端的设备，这样在每个链路两端控制传输(发送和接收)的特性就被很好地定义和控制了。

InfiniBand通过交换机在节点之间直接创建一个私有的、受保护的通道，进行数据和消息的传输，无需CPU参与远程直接内存访问(RDMA)和发送/接收由InfiniBand适配器管理和执行的负载。

适配器通过PCI Express接口一端连接到CPU，另一端通过InfiniBand网络端口连接到InfiniBand子网。与其他网络通信协议相比，这提供了明显的优势，包括更高的带宽、更低的延迟和增强的可伸缩性。

1. **Infiniband架构**

InfiniBand Architecture(IBA)是为硬件实现而设计的，而TCP则是为软件实现而设计的。因此，InfiniBand是比TCP更轻的传输服务，因为它不需要重新排序数据包，因为较低的链路层提供有序的数据包交付。传输层只需要检查包序列并按顺序发送包。

进一步，因为InfiniBand提供以信用为基础的流控制(发送方节点不给接收方发送超出广播 “信用“大小的数据包),传输层不需要像TCP窗口算法那样的包机制确定最优飞行包的数量。这使得高效的产品能够以非常低的延迟和可忽略的CPU利用率向应用程序交付56、100Gb/s的数据速率。

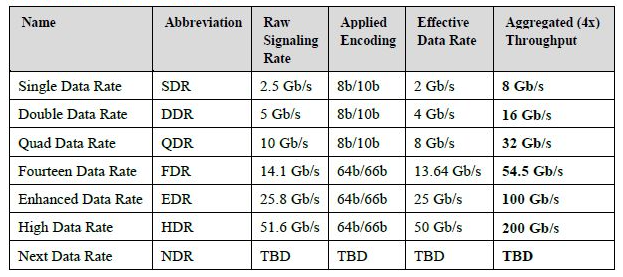
IB是以通道(Channel)为基础的双向、串行式传输，在连接拓朴中是采用交换、切换式结构(Switched Fabric)，所以会有所谓的IBA交换器(Switch)，此外在线路不够长时可用IBA中继器(Repeater)进行延伸。

而每一个IBA网络称为子网(Subnet)，每个子网内最高可有65,536个节点(Node)，IBASwitch、IBA Repeater仅适用于Subnet范畴，若要通跨多个IBA Subnet就需要用到IBA路由器(Router)或IBA网关器(Gateway)。至于节点部分，Node想与IBA Subnet接轨必须透过配接器(Adapter)，若是CPU、内存部分要透过HCA (Host Channel Adapter)，若为硬盘、I/O部分则要透过TCA (Target Channel Adapter)，之后各部分的衔接称为联机(Link)。上述种种构成了一个完整的IBA。

1. **InfiniBand速率发展**

InfiniBand串行链路可以在不同的信令速率下运行，然后可以捆绑在一起实现更高的吞吐量。原始信令速率与编码方案耦合，产生有效的传输速率。编码将通过铜线或光纤发送的数据的错误率降至最低，但也增加了一些开销(例如，每8位数据传输10位)。

典型的实现是聚合四个链接单元(4X)。目前，InfiniBand系统提供以下吞吐量速率:



InfiniBand的基本带宽是2.5Gb/s，这是InfiniBand 1.x。InfiniBand是全双工的，因此在两个方向上的理论最大带宽都是2.5Gb/s，总计5Gb/s。与此相反，PCI是半双工，因此32位、33MHz的PCI总线单个方向上能达到的理论最大带宽是1Gb/s，64位、133MHz的PCI-X总线能达到8.5Gb/s，仍然是半双工。当然，任何一种总线的实际吞吐量从来没有达到理论最大值。

如果要获取比InfiniBand 1.x更多的带宽，只要增加更多缆线就行。InfiniBand 1.0规范于2000年10月完成，支持一个通道内多个连接的网络，数据速率可提高4倍(10Gb/s)和12倍(30Gb/s)，也是双向的。

　　InfiniBand是在串行链路上实现超高速率的，因此电缆和连接器相对并行I/O接口PCI、IDE/ATA、SCSI和IEEE-1284来说，接口小也便宜。并行链路有一个固有的优势，因为它的多个缆线相当于高速公路上的多个车道，但现代的I/O收发器芯片使串行链路达到更高的数据速率，并且价格便宜。这就是为什麽最新的技术――InfiniBand、IEEE-[1394](https://product.pconline.com.cn/itbk/diy/mb/1112/2628159.html" \t "https://product.pconline.com.cn/itbk/life/electronic/1802/_blank)、串行ATA、串行连接SCSI、USB采用串行I/O而不是并行I/O。

1. **summit性能分析**

连接两种不同的I/O标准通常要增加数据通道的延迟。在最坏的情况下，将InfiniBand网络引入到一个已经安装多个不同技术组成的网络中，如果组织管理差，会降低其整体性能。InfiniBand的支持者声称理想的解决方案是完整的InfiniBand体系结构。任何部件都可以直接连接到InfiniBand网络，可以使用优化的文件协议，最好是使用直接访问文件系统(DAFS)。

　　DAFS独立于传输，是基于NFS的共享式文件访问协议。它是优化过的，用于1到100台机器的集群服务器环境中的I/O密集、CPU受限、面向文件的任务。典型的应用包括数据库、web服务、e-mail和地理信息系统(GIS)，当然也包括存储应用。