Infiniband 网络架构分析

·介绍

InfiniBand（直译为“无限带宽”技术，缩写为 **IB**）是一个用于高性能计算的计算机网络通信标准，它具有极高的吞吐量和极低的延迟，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。Infiniband 开放标准技术简化并加速了服务器之间的连接,同时支持服务器与远程存储和网络设备的连接。

·发展历程

1999 年开始起草规格及标准规范，2000 年正式发表，但发展速度不及 Rapid I/O、PCI-X、PCI-E 和 FC，加上 Ethernet 从 1Gbps 进展至 10Gbps。所以直到 2005 年之后，InfiniBand Architecture(IBA)才在集群式超级计算机上广泛应用。全球 Top 500 大效能的超级计算机中有相当多套系统都使用上 IBA。随着越来越多的大厂商正在加入或者重返到它的阵营中来，包括 Cisco、IBM、HP、Sun、NEC、Intel、LSI 等。InfiniBand 已经成为目前主流的高性能计算机互连技术之一。为了满足 HPC、企业数据中心和云计算环境中的高 I/O 吞吐需求，新一代高速率 56Gbps 的 FDR (Fourteen Data Rate) 和 EDR InfiniBand 技术已经出现。

·IB 的基本概念

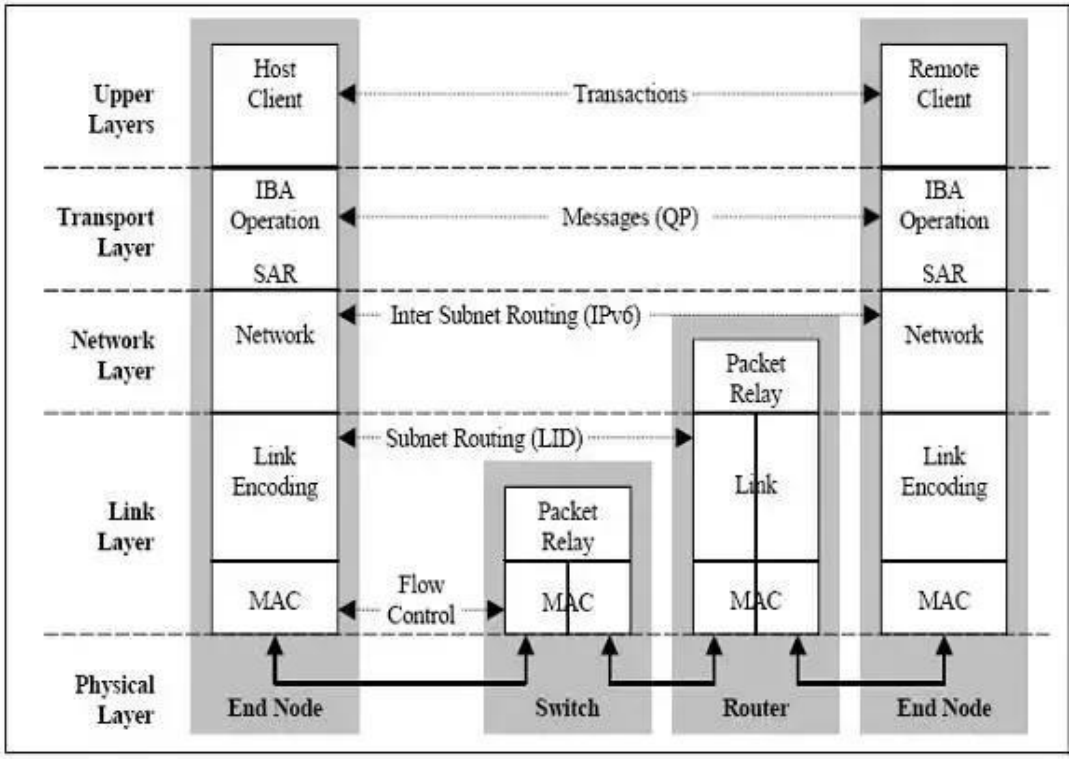
IB 是以通道为基础的双向、串行式传输，在连接拓朴中是采用交换、切换式结构(Switched Fabric)，在线路不够长时可用 IBA 中继器(Repeater)进行延伸。每一个 IBA 网络称为子网(Subnet)，每个子网内最高可有 65,536 个节点(Node)，IBA Switch、IBARepeater 仅适用于 Subnet 范畴，若要通跨多个 IBASubnet 就需要用到 IBA 路由器(Router)或 IBA 网关器(Gateway)。

每个节点(Node) 必须透过配接器(Adapter)与 IBA Subnet 连接，节点 CPU、内存要透过 HCA(Host Channel Adapter)连接到子网；节点硬盘、I/O 则要透过 TCA(TargetChannel Adapter)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(Backplane 背板)，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。

·IB 协议简介

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)。

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。

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。

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。

·应用案例

应用案例 1：大数据持续爆发

像 Oracle Exadata 这样的大数据方案提供商已经将 InfiniBand 部署在他们的设备内部好几年了，对于需要高速计算能力和大量内部数据流的横向扩展应用程序来说，InfiniBand 是一个理想的交换结构解决方案。hadoop 集群中的 InfiniBand 网络将分析吞吐量翻了一倍，因为需要更少的计算节点，由此累计省下来的开支比部署 InfiniBand 网络通常要多很多。只要数据在容量和种类上持续增长，高性能的数据交换就会一直是数据中心网络的一个巨大挑战，如此一来高带宽和低延迟的 InfiniBand 网络使得它成为一个具有竞争力的替代方案。

应用案例 2：虚拟数据中心的虚拟 I/O

为了满足虚拟环境下关键应用程序的可用性和性能，IT 必须虚拟整个 I/O 的数据路径，包括共享存储和连接网络，反过来，I/O 数据路径，必须能够支持多协议和动态配置。

为了加强运行关键程序的虚拟机的移动性，虚拟机必须能够无缝而且快速的将整个网络和存储迁移到另外一个地方。这意味着支撑其运作的物理架构必须被每台主机同等的连接到。但是由于主机一般都有不同的物理网卡和物理网络连接，虚拟机的无缝移动成了一个挑战。像 InfiniBand 这样聚合、平坦的网络架构提供了一个巨大的管道，可以动态按需分配，这使得其在紧密、高移动性的虚拟环境下变得尤为理想。