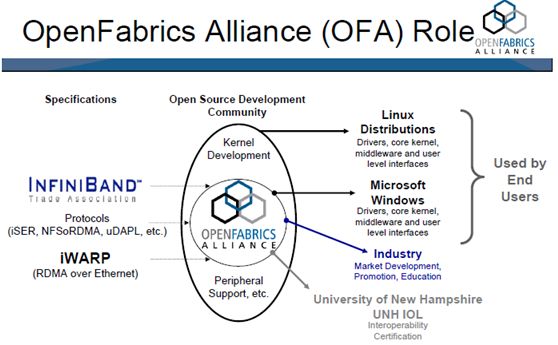
**Infiniband网络结构分析**

**Infiniband技术前言**

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。OpenFabrics Enterprise Distribution (OFED)是一组开源软件驱动、核心内核代码、中间件和支持InfiniBand Fabric的用户级接口程序。

2005年由OpenFabrics  Alliance (OFA)发布第一个版本。Mellanox OFED用于Linux，Windows (WinOF)，包括各种诊断和性能工具，用于监视InfiniBand网络的运行情况，包括监视传输带宽和监视Fabric内部的拥塞情况。

OpenFabrics Alliance (OFA)是一个基于开源的组织，它开发、测试、支持OpenFabrics企业发行版。该联盟的任务是开发并推广软件，通过将高效消息、低延迟和最大带宽技术架构直接应用到最小CPU开销的应用程序中，从而实现最大应用效率。



该联盟成立于2004年6月，最初是OpenIB联盟，致力于开发独立于供应商、基于Linux的InfiniBand软件栈。2005，联盟致力于支持Windows，此举将使软件栈真正跨平台。

2006年，该组织再次扩展其章程，包括对iWARP的支持，在2010年增加了对RoCE (RDMA over Converged)支持通过以太网交付高性能RDMA和内核旁路解决方案。2014年，随着OpenFabrics Interfaces工作组的建立，联盟再次扩大，实现对其他高性能网络的支持

**IB技术的发展**

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HPC高算系统TOP500大效能的超级计算机中有相当多套系统都使用上IBA。

除了InfiniBand Trade Association (IBTA)9个主要董事成员CRAY、Emulex、HP、IBM、intel、Mellanox、Microsoft、Oracle、Qlogic专门在应用和推广InfiniBand外，其他厂商正在加入或者重返到它的阵营中来，包括Cisco、Sun、NEC、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100Gpb EDR InfiniBand技术已经广泛应用。

**IB技术的优势**

Infiniband大量用于FC/IP SAN、NAS和服务器之间的连接,作为iSCSI RDMA的存储协议iSER已被IETF标准化。目前EMC全系产品已经切换到Infiniband组网，IBM/TMS的FlashSystem系列，IBM的存储系统XIV Gen3，DDN的SFA系列都采用Infiniband网络。

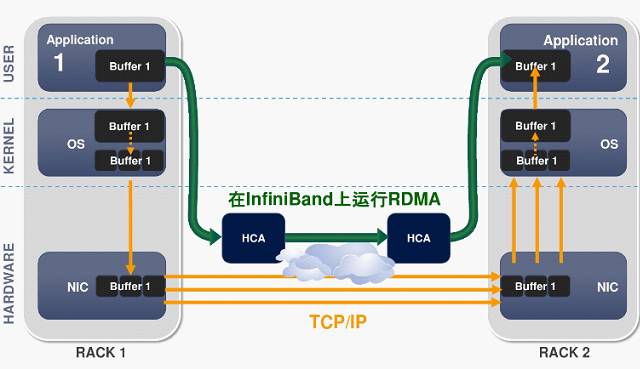
相比FC的优势主要体现在性能是FC的3.5倍，Infiniband交换机的延迟是FC交换机的1/10，支持SAN和NAS。

存储系统已不能满足于传统的FC SAN所提供的服务器与裸存储的网络连接架构。HP SFS和IBM GPFS 是在Infiniband fabric连接起来的服务器和iSER Infiniband存储构建的并行文件系统，完全突破系统的性能瓶颈。

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)。

RDMA技术实现内核旁路，可以提供远程节点间RDMA读写访问，完全卸载CPU工作负载，基于硬件传出协议实现可靠传输和更高性能。



相比TCP/IP网络协议，IB使用基于信任的、流控制的机制来确保连接的完整性，数据包极少丢失，接受方在数据传输完毕之后，返回信号来标示缓存空间的可用性，所以IB协议消除了由于原数据包丢失而带来的重发延迟，从而提升了效率和整体性能。

TCP/IP具有转发损失的数据包的能力，但是由于要不断地确认与重发，基于这些协议的通信也会因此变慢，极大地影响了性能

**IB基本概念**

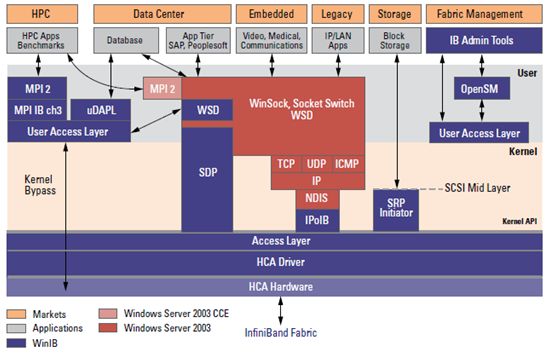
IB是以通道为基础的双向、串行式传输，在连接拓朴中是采用交换、切换式结构(Switched Fabric)，在线路不够长时可用IBA中继器(Repeater)进行延伸。每一个IBA网络称为子网(Subnet)，每个子网内最高可有65,536个节点(Node)，IBA Switch、IBARepeater仅适用于Subnet范畴，若要通跨多个IBASubnet就需要用到IBA路由器(Router)或IBA网关器(Gateway)。

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。

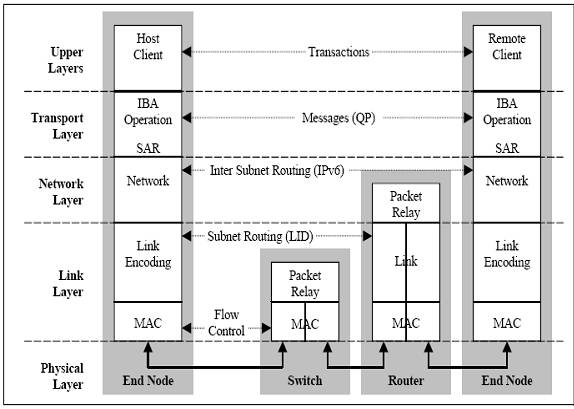
**IB协议简介**

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。



Mellanox OFED是一个单一的软件堆栈，包括驱动、中间件、用户接口，以及一系列的标准协议IPoIB、SDP、SRP、iSER、RDS、DAPL(Direct Access Programming Library)，支持MPI、Lustre/NFS over RDMA等协议，并提供Verbs编程接口；Mellanox OFED由开源OpenFabrics组织维护。

当然，Mellanox OFED软件堆栈是承载在InfiniBand硬件和协议之上的，软件通协议和硬件进行有效的数据传输。



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。

**IB基本应用场景**

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)

