Infiniband 网络结构分析

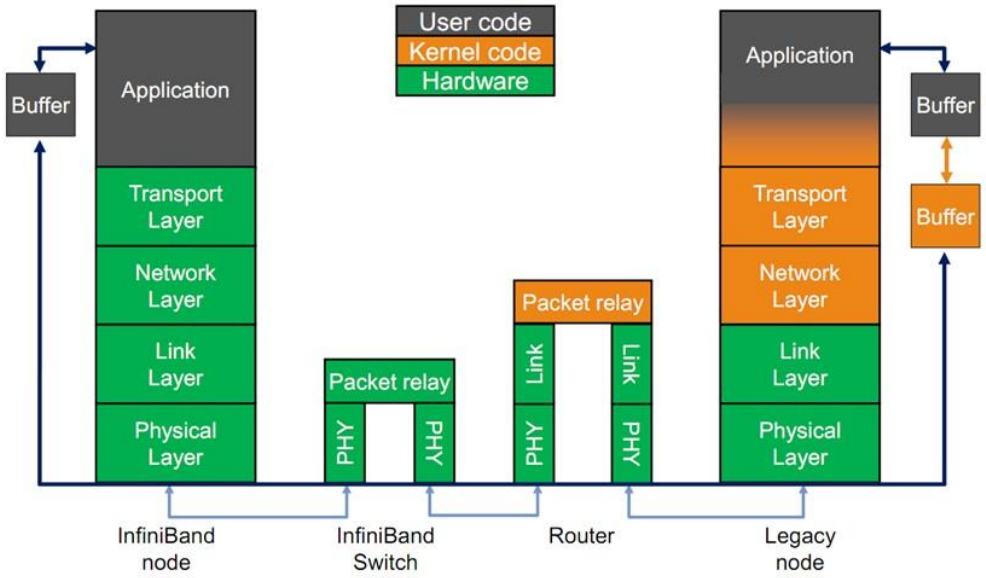
一、Infiniband 网络简介

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。InfiniBand 最初目标是把 PCI 总线网络化，所以 InfiniBand 除了具有很强的网络性能以外还直接继承了总线的高带宽和低时延。其支持的可寻址设备高达 64000 个。

InfiniBand 与通过以太网基础设施（NIC、交换机和路由器）运行并支持传统（基于 IP）应用程序不同。InfiniBand 作为一种完全不同的协议，使用不同的寻址机制，它不是 IP，并且不支持套接字，因此它不支持旧版应用程序。这意味着在某些群集中，InfiniBand 不能用作唯一的互连，因为许多管理系统都是基于 IP 的。其结果可能是，使用 InfiniBand 进行数据流量的群集也可能在群集中部署以太网基础结构（用于管理）。这增加了群集部署的价格和复杂性。

二、Infiniband 网络协议层次与网络结构

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。



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(Queue Pair, 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。

物理层还定义了用于形成帧的符号(包的开始和结束)、数据符号(Data Symbols)、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(Sockets Direct 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。

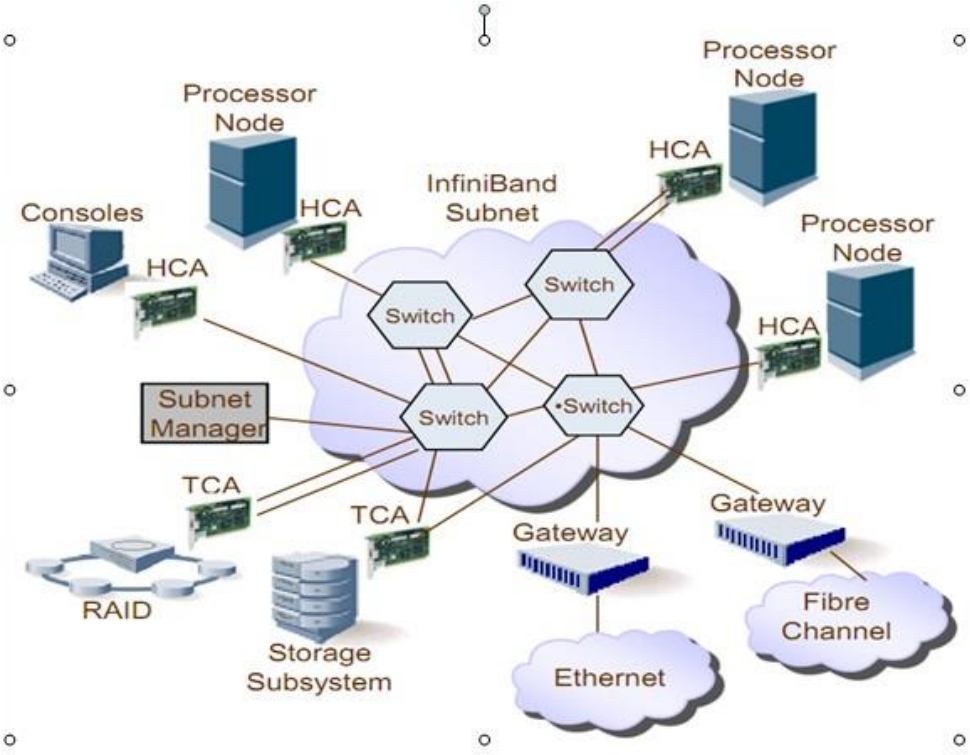
iSER (iSCSIRDMA Protocol)类似于 SRP(SCSI RDMA protocol)协议，是 IB SAN 的一种协议 ，其主要作用是把 iSCSI 协议的命令和数据通过 RDMA 的方式跑到例如 Infiniband 这种网络上，作为 iSCSI RDMA 的存储协议 iSER 已被 IETF 所标准化。

RDS(Reliable Datagram 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。

Infiniband 的网络拓扑结构如下图，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；

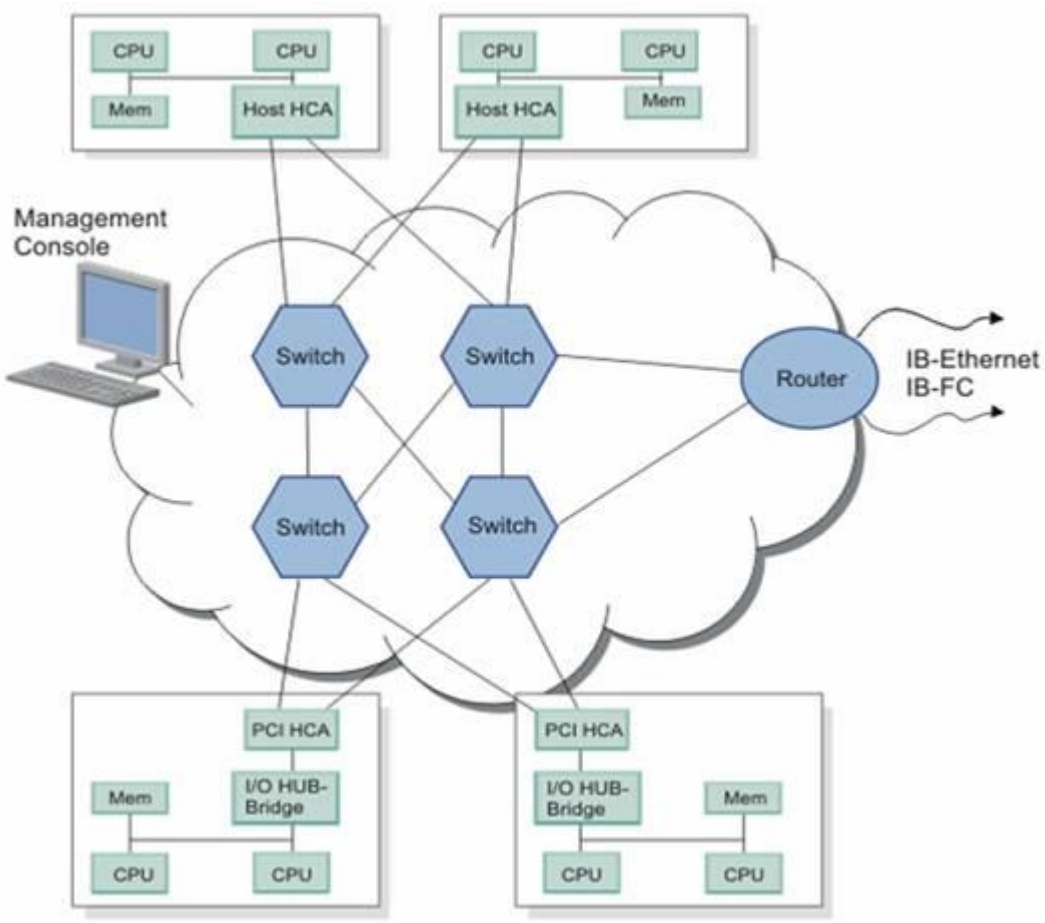
（2）TCA(Target Channel Adapter)，它将 I/O 设备（例如网卡、SCSI 控制器）的数字信号打包发送给 HCA；

（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；

三、InfiniBand 系统网络

InfiniBand 系统网络主要由两个核心部件组成：主机通道适配器（Host Channel Adapter）和 InfiniBand 交换机。其中，HCA 为主机设备提供一个接口用于支持所有 InfiniBand 定义的操作，而交换机则用于将一个端口接收到的 InfiniBand 报文转发到另一个端口，它支持单播和多播两种机制。下图为网络结构图：


CPU
CPU
Mem
Host HCA
CPU
CPU
Host HCA
Mem
Management
Console
Switch
Switch
Switch
Switch
Router
IB-Ethernet
IB-FC
PCI HCA
I/O HUB-
Bridge
Mem
CPU
CPU
PCI HCA
I/O HUB-
Bridge
Mem
CPU
CPU