Infiniband 网络结构分析

201708010814-李宗鸿

一、 简介

1、 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。

2、与目前计算机的 I/O 子系统不同，InfiniBand 是一个功能完善的网络通信系统。InfiniBand 贸易组织把这种新的总线结构称为 I/O 网络，并把它比作开关，因为所给信息寻求其目的地址的路径是由控制校正信息决定的。InfiniBand 使用的是网际协议版本 6 的 128 位地址空间，因此它能提供近乎无限量的设备扩展性。

3、通过 InfiniBand 传送数据时，数据是以数据包方式传输，这些数据包会组合成一条条信息。这些信息的操作方式可能是远程直接内存存取的读写程序，或者是通过信道接受发送的信息，或者是多点传送传输。就像大型机用户所熟悉的信道传输模式，所有的数据传输都是通过信道适配器来开始和结束的。每个处理器（例如个人电脑或数据中心服务器）都有一个主机通道适配器，而每个周边设备都有一个目标通道适配器。通过这些适配器交流信息可以确保在一定服务品质等级下信息能够得到有效可靠的传送。

二、 基本概念

1、 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)。

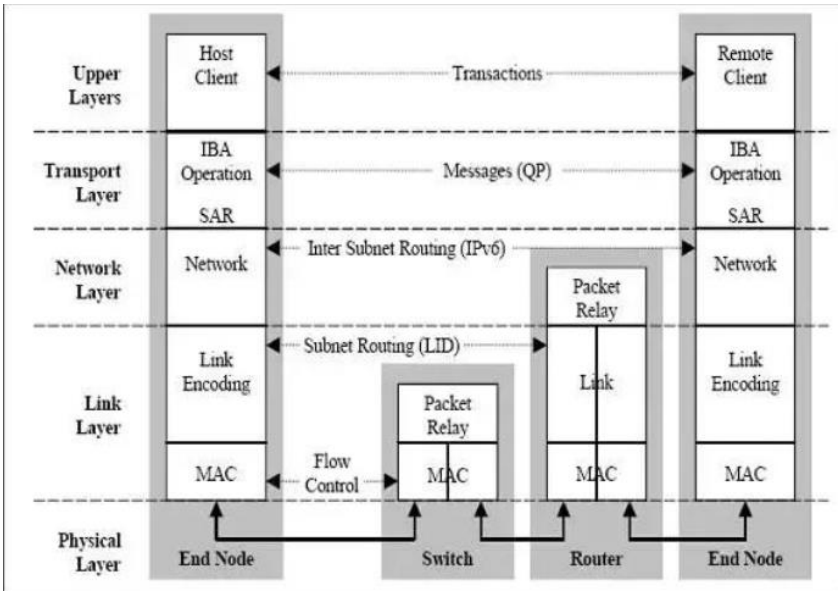
2、 每个节点(Node) 必须透过配接器(Adapter)与 IBA Subnet 连接，节点 CPU、内存要透过 HCA(Host Channel Adapter)连接到子网；节点硬盘、I/O 则要透过 TCA(TargetChannel Adapter)连接到子网，这样的一个拓扑结构就构成了一个完整的 IBA。

3、 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。

三、 协议层次

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，如何将比特信号组 成符号，然后再组成帧)、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

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。

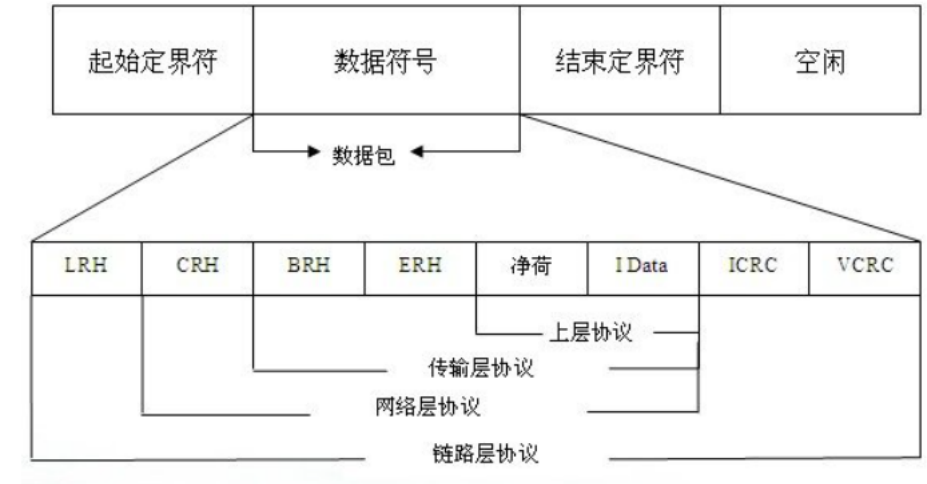
接收端的队列负责将数据重组到指定的数据缓冲区中。除了原始数据报外，所有的数据包都包含 BTH，BTH 指定目的队列并指明操作类型、数据包序列号和分区信息。

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(SocketsDirect 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(ReliableDatagram 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。

数据包格式：

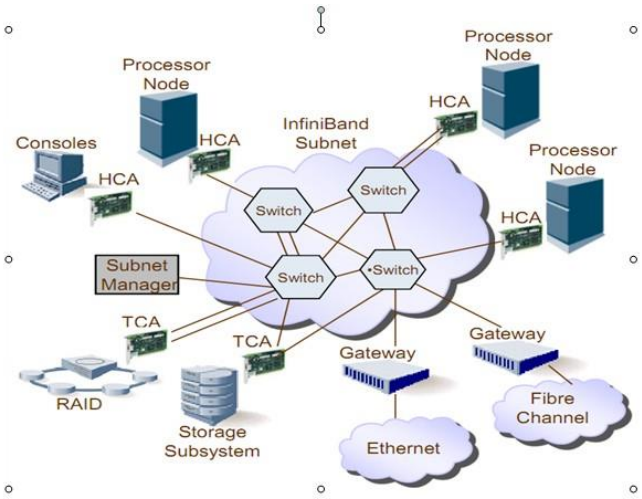


LRH：本地路由报头
BRH：基本传送报头
IData：即时数据
环冗余校验
VCRC：可变循环冗余校验

CRH：全局路由报头
ERH：扩展传送报头
ICRC：恒定循

四、 网络结构

Infiniband 的网络拓扑结构如图，其组成单元主要分为以下几类



（1） HCA（Host Channel Adapter）主机通道适配器：IB 连接的服务器的网络接口，提供虚拟/物理内存的映像，内存直接访问和内存保护。并提供 RDMA 传送数据。HCA 主要功能就是用硬件设备实现了高效的通信功能。

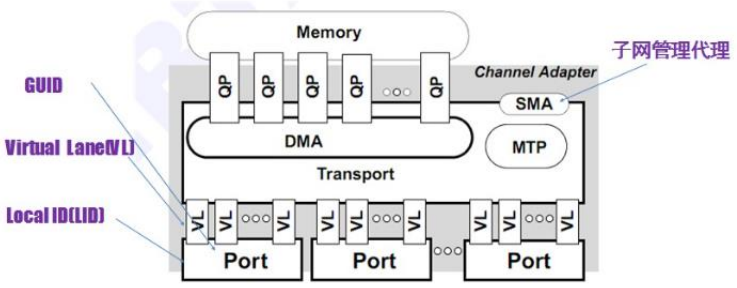
（2） TCA(Target Channel Adapter) 目标通道适配器：与 HCA 类似，但不需要虚拟内存和映像。提供到 I/O 控制器的链路和传输服务，使 I/O 设备可脱离主机而直接置于网络中，实现了处理计算、存储 I/O、网络 I/O 等功能的独立。它将 I/O 设备（例如网卡、SCSI 控制器）的数字信号打包发送给 HCA；

（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） Switch：IBA 中提供高集中带宽、负载平衡等的关键部件。一个基本的 Switch 芯片支持 24 端口或 36 端口，可以提供构建成千上万个全交换的高效双向带宽的网络端口，提供无阻塞、低延迟交换功能。

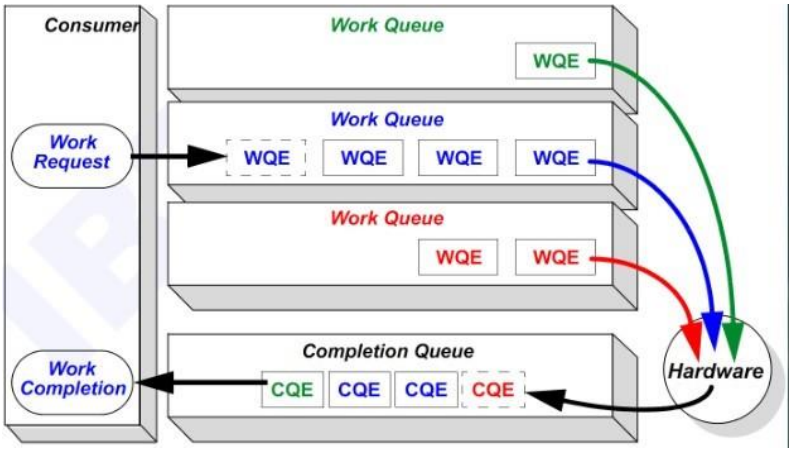
（5） SM（Subnet Manager）子网管理器：IB 网络叫做 InfiniBand 子网，所有的设备都在一个 SM 下控制。负责配置和管理 Switch、路由器及通道适配器的应用程序。

（6）交换机和路由器；

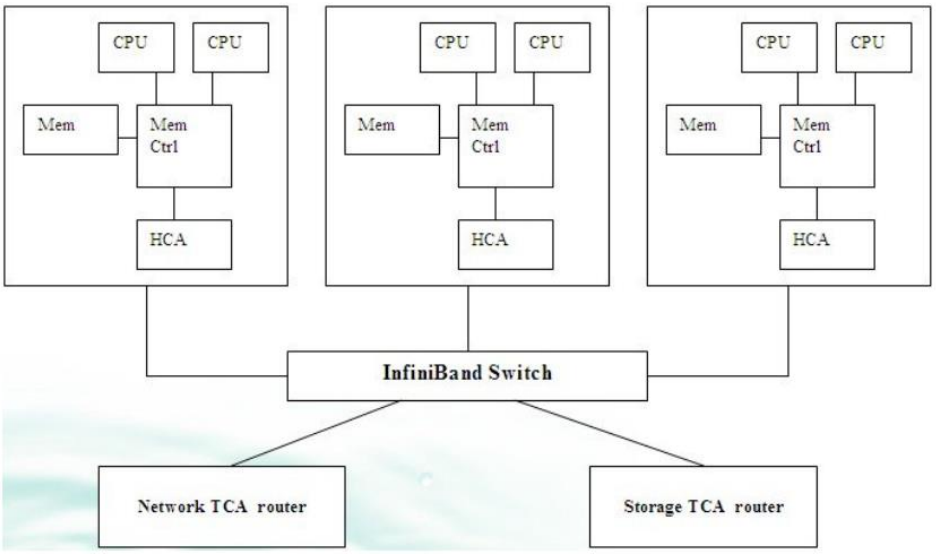


如图所示每个端口具有一个 GUID(Globally Unique Identifier)，GUID 是全局唯一的，类似于以太网 MAC 地址。运行过程中，子网管理代理（SMA）会给端口分配一个本地标识（LID），LID 仅在子网内部有用。

QP 是 infiniband 的一个重要概念，它是指发送队列和接收队列的组合，用户调用 API 发送接收数据的时候，实际上是将数据放入 QP 当中，然后以轮询的方式将 QP 中的请求一条条的处理，其模式类似于生产者-消费者模式。



InfiniBand 互联结构



InfiniBan 将 I/O 系统与复杂的 CPU/存储器分开，采用基于通道的高速串行链路和可扩

展的光纤交换网络替代共享总线结构，提供了高带宽,低延迟,可扩展 的 I/O 互连，克服了传统的共享 I/O 总线结构的种种弊端。