湖南大學

HUNAN UNIVERSITY

计算机系统设计

选题名称：Infiniband 网络架构分析

姓　　名：肖若愚

学　　号：201708010619

专业班级：物联 1702

一、Infiniband 介绍

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。由于它具有高带宽、低延时、 高可扩展性的特点，它非常适用于服务器与服务器（比如复制，分布式工作等），服务器和存储设备（比如 SAN 和直接存储附件）以及服务器和网络之间（比如 LAN， WANs 和 the Internet）的通信 。

随着 CPU 性能的飞速发展，I/O 系统的性能成为制约服务器性能的瓶颈。于是人们开始重新审视使用了十几年的 PCI 总线架构。虽然 PCI 总线结构把数据的传输从 8 位/16 位一举提升到 32 位，甚至当前的 64 位，但是它的一些先天劣势限制了其继续发展的势头。PCI 总线有如下缺陷：

(1)由于采用了基于总线的共享传输模式，在 PCI 总线上不可能同时传送两组以上的数据，当一个 PCI 设备占用总线时，其他设备只能等待；

(2)随着总线频率从 33MHz 提高到 66MHz，甚至 133MHz（PCI-X），信号线之间的相互干扰变得越来越严重，在一块主板上布设多条总线的难度也就越来越大；

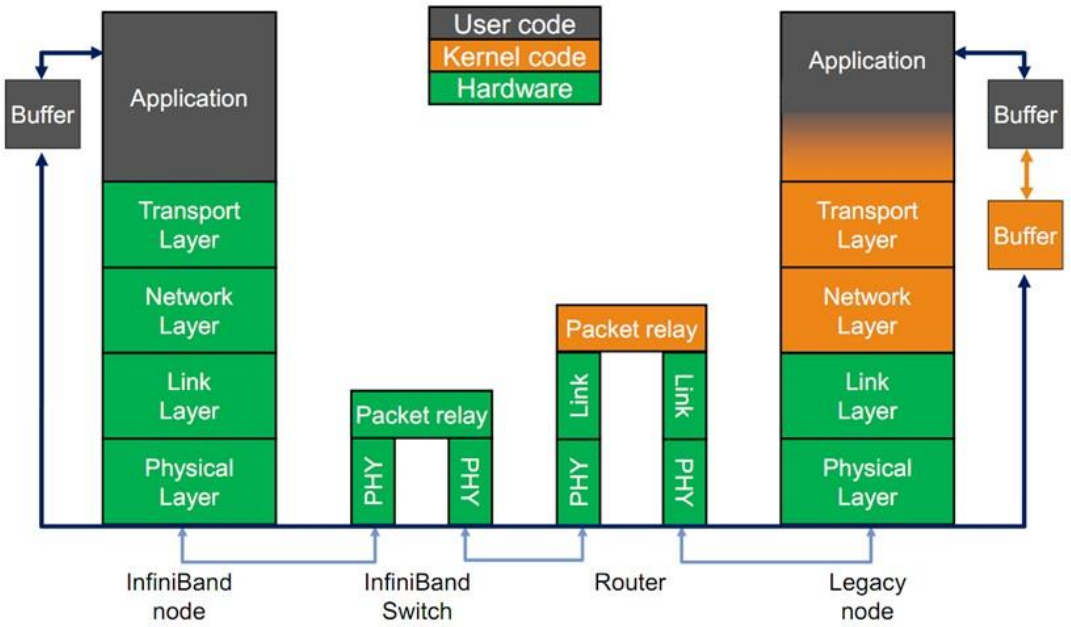
(3)由于 PCI 设备采用了内存映射 I/O 地址的方式建立与内存的联系，热添加 PCI 设备变成了一件非常困难的工作。目前的做法是在内存中为每一个 PCI 设备划出一块 50M 到 100M 的区域，这段空间用户是不能使用的，因此如果一块主板上支持的热插拔 PCI 接口越多，用户损失的内存就越多；

(4)PCI 的总线上虽然有 buffer 作为数据的缓冲区，但是它不具备纠错的功能，如果在传输的过程中发生了数据丢失或损坏的情况，控制器只能触发一个 NMI 中断通知操作系统在 PCI 总线上发生了错误

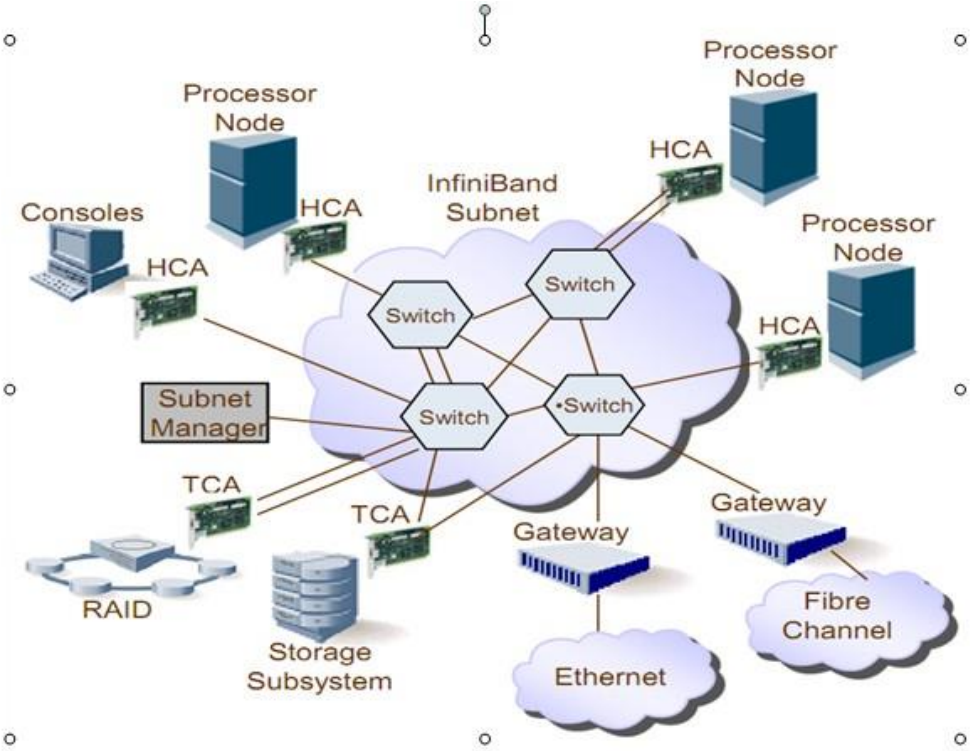
因此，Intel、 Cisco、 Compaq、 EMC、 富士通等公司共同发起了 infiniband 架构，其目的是为了取代 PCI 成为系统互连的新技术标准，其核心就是将 I/O 系统从服务器主机中分离出来。

InfiniBand 采 用双队列程序提取技术,使应用程序直接将数据从适配器 送入到应用内存(称为远程直接存储器存取或 RDMA), 反之依然。在 TCP/IP 协议中,来自网卡的数据先拷贝到 核心内存,然后再拷贝到应用存储空间,或从应用空间 将数据拷贝到核心内存,再经由网卡发送到 Internet。这 种 I/O 操作方式,始终需要经过核心内存的转换,它不 仅增加了数据流传输路径的长度,而且大大降低了 I/O 的访问速度,增加了 CPU 的负担。而 SDP 则是将来自网 卡的数据直接拷贝到用户的应用空间,从而避免了核心 内存参入。这种方式就称为零拷贝,它可以在进行大量 数据处理时,达到该协议所能达到的最大的吞吐量

二、Infiniband 的协议层次和网络结构



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(QueuePair,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。



Infiniband 的网络拓扑结构如图 2，其组成单元主要分为四类：

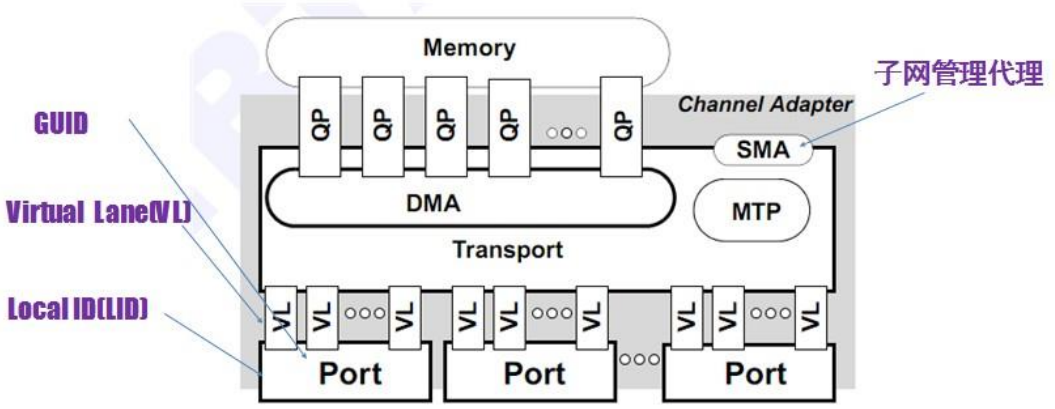
（1）HCA（Host Channel Adapter），它是连接内存控制器和 TCA 的桥梁；

（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；

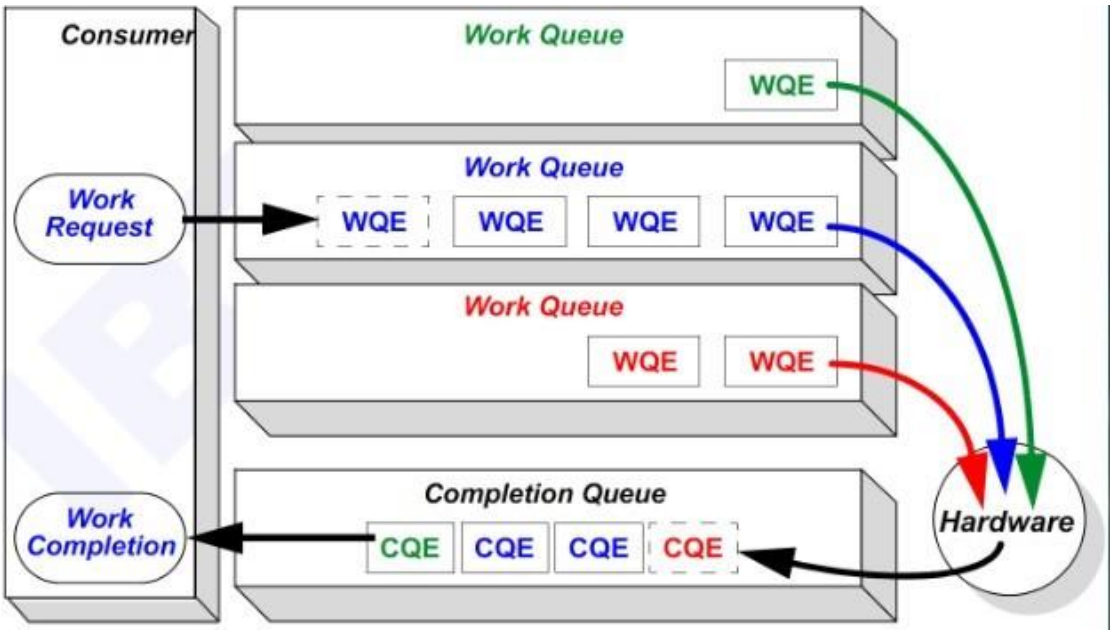
（3）Infiniband link，它是连接 HCA 和 TCA 的光纤，InfiniBand 架构允许硬件厂家以 1 条、4 条、12 条光纤 3 种方式连结 TCA 和 HCA；

（4）交换机和路由器；

无论是 HCA 还是 TCA，其实质都是一个主机适配器，它是一个具备一定保护功能的可编程 DMA（Direct Memory Access，直接内存存取 ）引擎，



如图 3 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。



如图 4 所示，图中 Work queue 即是 QP 中的 send Queue 或者 receive Queue，WQ 中的请求被处理完成之后，就被放到 Work Completion 中。