Infiniband 网络结构分析

一、概述

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。

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。

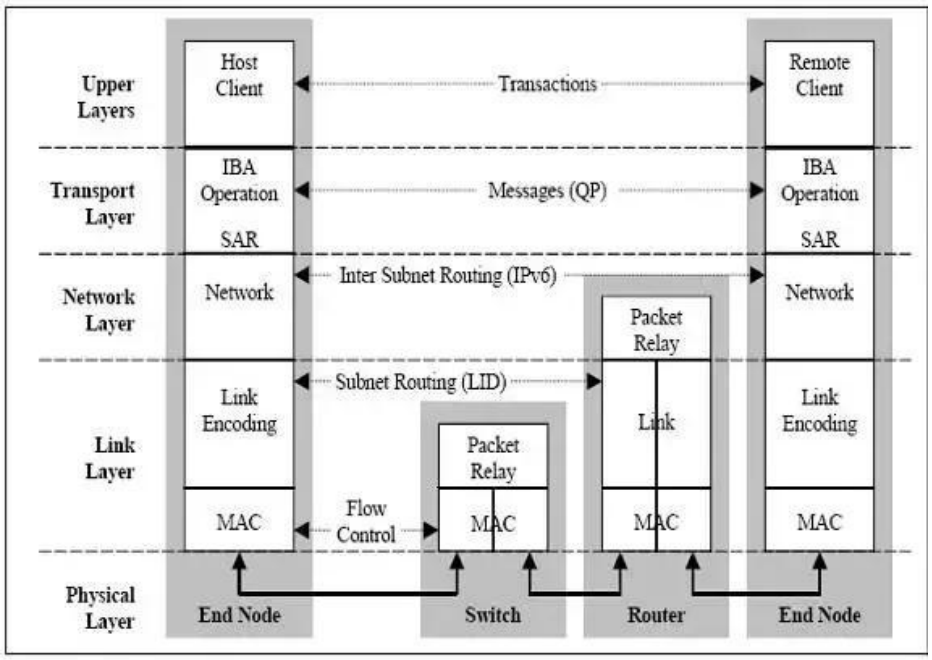
通过 InfiniBand 传送数据时，数据是以数据包方式传输，这些数据包会组合成一条条信息。这些信息的操作方式可能是远程直接内存存取的读写程序，或者是通过信道接受发送的信息，或者是多点传送传输。就像大型机用户所熟悉的信道传输模式，所有的数据传输都是通过信道适配器来开始和结束的。每个处理器（例如个人电脑或数据中心服务器）都有一个主机通道适配器，而每个周边设备都有一个目标通道适配器。通过这些适配器交流信息可以确保在一定服务品质等级下信息能够得到有效可靠的传送。

二、IB 协议简介及网络结构

• IB 协议：

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。



各层简述：

• 物理层(Physical Layer)

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。

• 链路层(Link Layer)

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。

• 网络层(Network Layer)

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。

• 传输层(Transport Layer)

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。

• 上层协议(Upper Layers)

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。

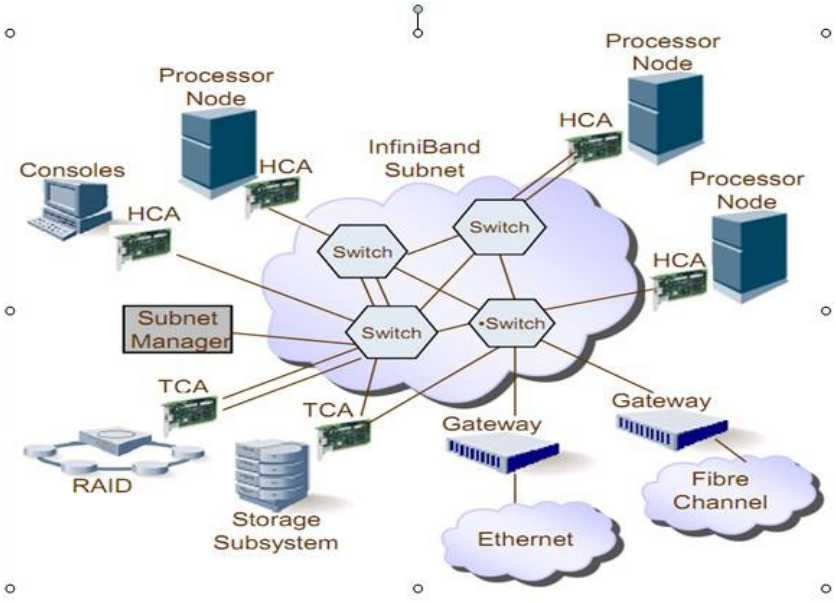
• iSER(iSCSIRDMA Protocol)类似于 SRP(SCSI RDMA protocol)协议，是 IB SAN 的一种协议 ，其主要作用是把 iSCSI 协议的命令和数据通过 RDMA 的方式跑到例如 Infiniband 这种网络上，作为 iSCSI RDMA 的存储协议 iSER 已被 IETF 所标准化。

• 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。

• IB 网络拓扑结构：



Infiniband 的网络拓扑结构如图所示，其组成单元主要分为四类：

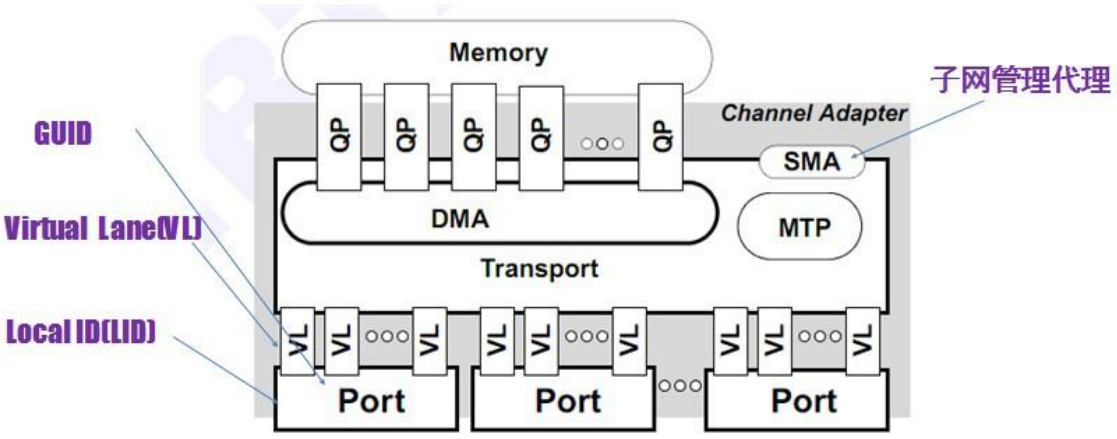
• HCA（Host Channel Adapter），它是连接内存控制器和 TCA 的桥梁；

• 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；

• 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；

• 交换机和路由器；

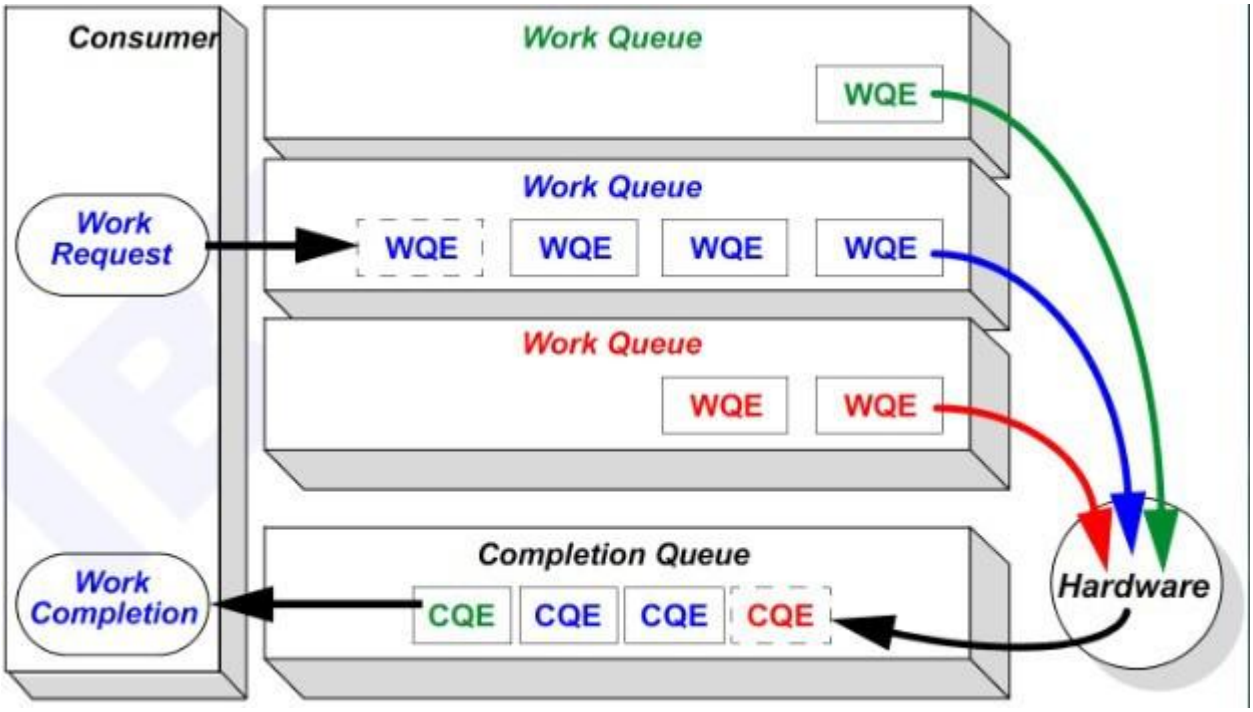
无论是 HCA 还是 TCA，其实质都是一个主机适配器，它是一个具备一定保护功能的可编程 DMA（Direct Memory Access，直接内存存取 ）引擎，如图所示：



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

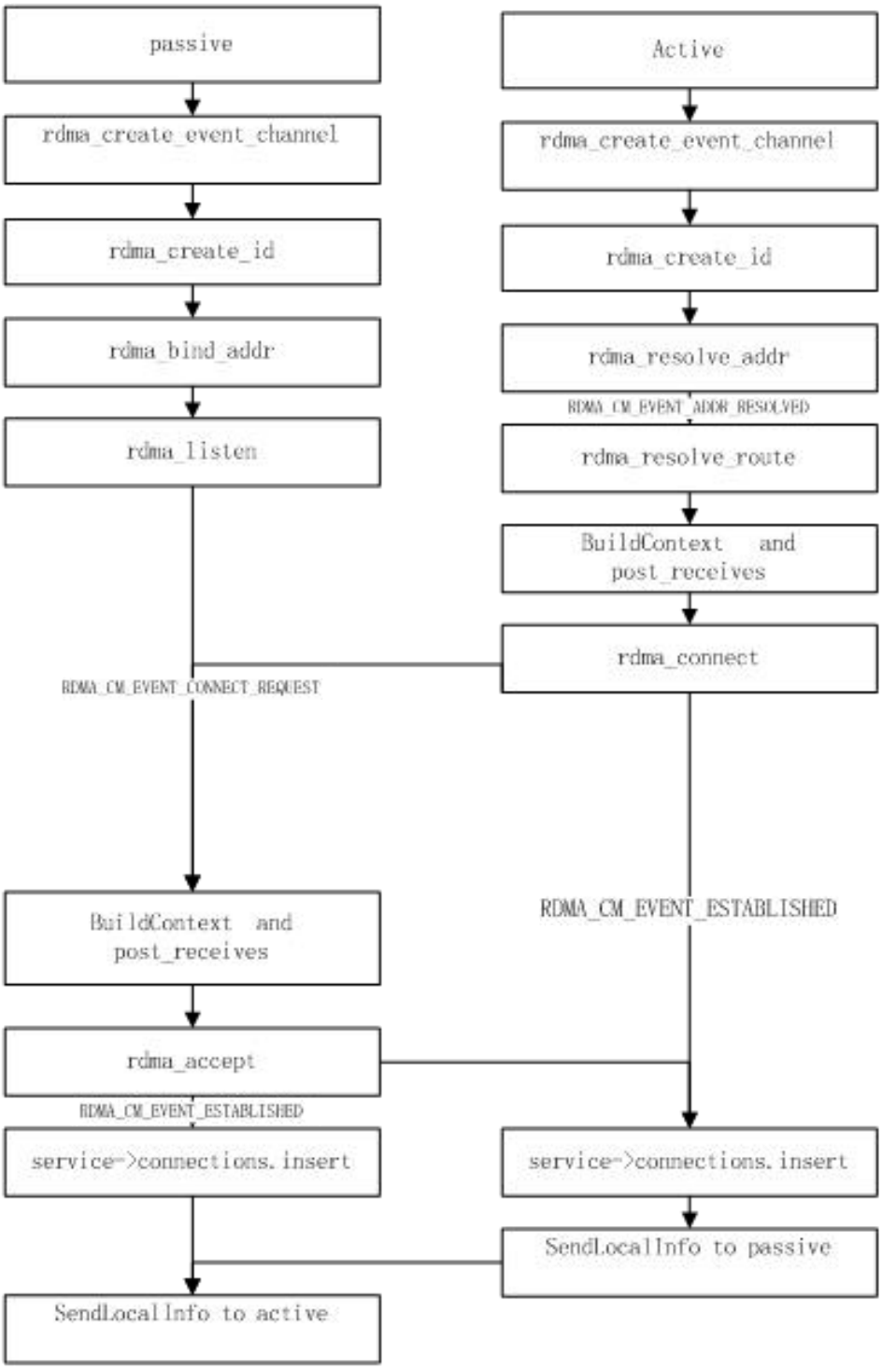
户调用 API 发送接收数据的时候，实际上是将数据放入 QP 当中，然后以轮询的方式将 QP 中的请求一条条的处理，其模式类似于生产者-消费者模式：



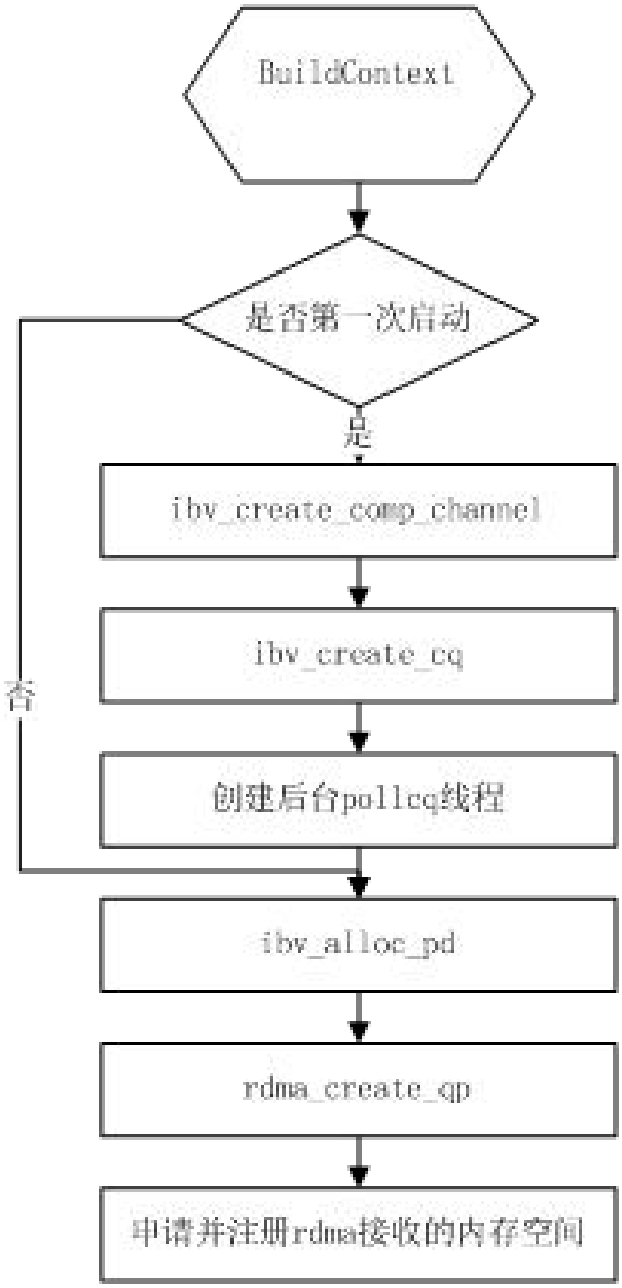
图中 Work queue 即是 QP 中的 send Queue 或者 receive Queue，WQ 中的请求被处理完成之后，就被放到 Work Completion 中。

三、IB 网络传送数据方式

Infiniband 提供了 VPI verbs API 和 RDMA_CM verbs API 这两个 API 集合，用户使用其中的库函数，就能很方便的在不同的机器之间传输数据。Infiniband 建立连接的流程如下图所示：



其中 buildcontext 的流程如下：



连接建立完成之后，就可以调用 ibv_post_recv 和 ibv_post_send 收发数据了，发送和接收请求都被放在 QP 中，后台需要调用 ibv_poll_cq 来逐条处理请求，由于 infiniband 连接中，一旦有一条数据发送或者接收失败，其后所有的数据发送或者接收都会失败，因此一旦检测到 WC 的状态不是成功，需要立即处理此错误（此时最好断开连接）。

常见错误：

ibv_poll_cq 处理完队列中的数据后，WC 会包含此次处理的全部信息，包括 wr_id、操作状态、错误码等等，错误码包含的信息对于解决错误非常有用：

- 错误码为 4（IBV_WC_LOC_PROT_ERR），这种错误通常意味着用户对内存的操作权限不够,需要检测在 ibv_post_recv 和 ibv_post_send 时 scatter/gather list 中传入的内存地址与长度是否正确，或者 ibv_reg_mr 操作是否成功。
- 错误码为 5，（IBV_WC_WR_FLUSH_ERR），在 flush 的时候出现错误，通常

是因为前一个操作出现了错误，接下来的一系列操作都会出现IBV_WC_WR_FLUSH_ERR 的错误。

• 错误码为 13（**IBV_WC_RNR_RETRY_EXC_ERR**），这种错误一般是因为本地 post 数据过快。在 infiniband 传输数据过程中，接收端首选需要注册内存并 ibv_post_recv 将此内存放入 receive queue 中然后发送端才能发送数据，如果接受端来不及完成这些操作发送端就发送数据，就会出现上述错误。

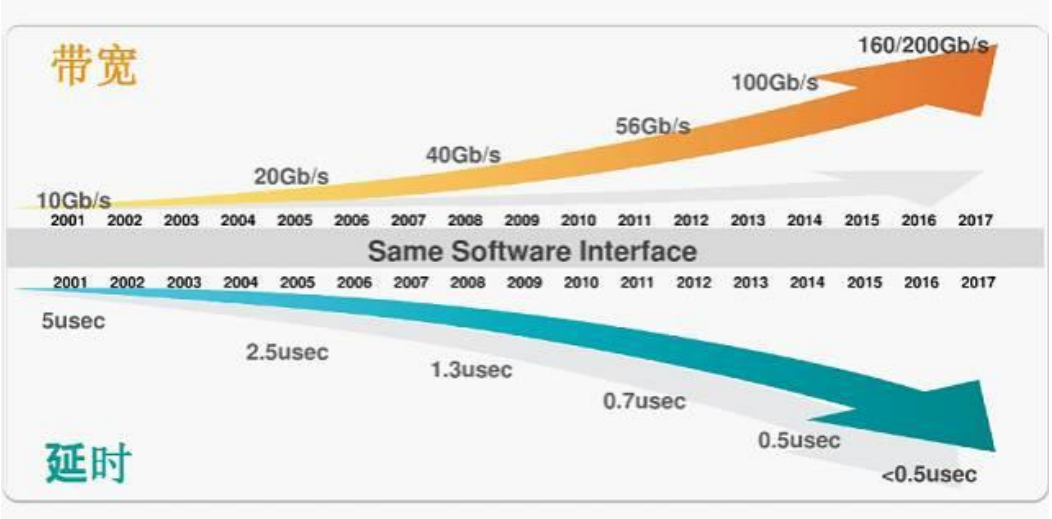
四、IB 技术的优点

•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。

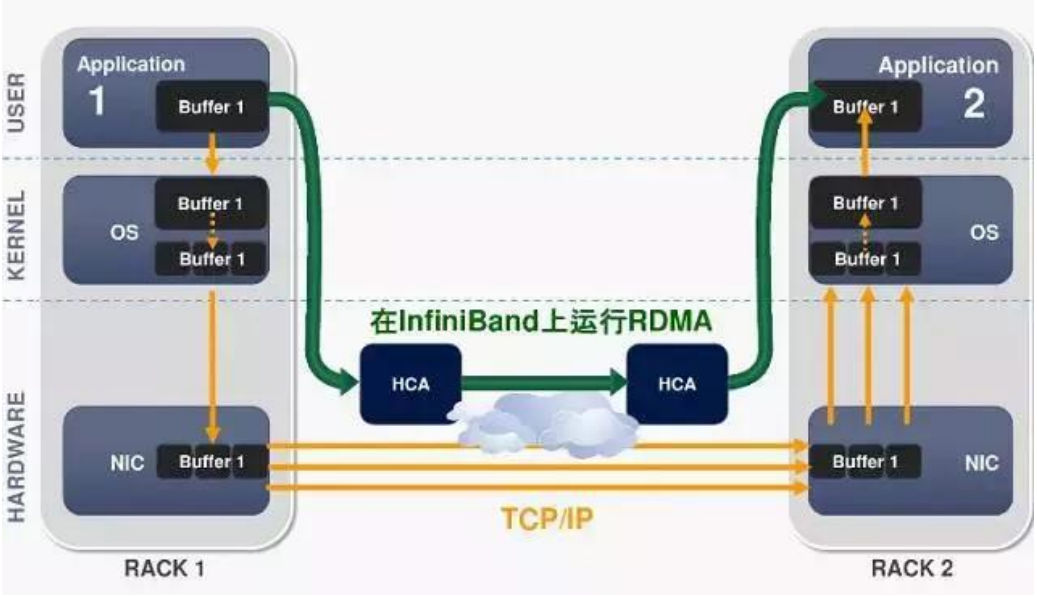
• 存储系统已不能满足于传统的 FC SAN 所提供的服务器与裸存储的网络连接架构。HP SFS 和 IBM GPFS 是在 Infiniband fabric 连接起来的服务器和 iSER Infiniband 存储构建的并行文件系统，完全突破系统的性能瓶颈。

•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 微秒、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)。



•IB 网络采用的 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。

参考资料：

• 百度百科-Infiniband：
https://baike.baidu.com/item/Infiniband/1963979?fr=aladdin
• Infiniband 技术和协议架构分析：
https://blog.csdn.net/swingwang/article/details/72887367
• Infiniband 学习总结：
https://www.cnblogs.com/D-Tec/p/3157582.html