

课程实验报告

课 程 名 称： 计算机系统设计

专 业 班 级： 物联1702

姓 名： 龚竞舟

学 号： 201708010612

完 成 时 间： 2019 年 12 月 7 日

计算机工程系

## 实验题目

Infiniband网络结构分析

**二．实验内容**

（1）Infiniband简介

Infiniband无限宽带技术（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联。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。

它的主要设计目的是针对服务器端的连接问题，将会被应用于服务器与服务器（比如复制、分布式工作），服务器与存储设备（比如SAN和直接存储附件）以及服务器和网络之间（比如LAN,WANs，THE INTERNET）的通信。与目前的计算机的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。

（2） Infiniband网络结构分层

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。

1. 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。
2. 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。
3. 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。
4. 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。
5. 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。
6. 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。

（3）Infiniband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) 。一方面Infiniband在主机侧采用RDMA技术释放CPU负载，可以把主机内数据处理的时延从几十微秒降低到1微秒；另一方面InfiniBand网络的高带宽(40G、56G和100G)、低时延(几百纳秒)和无丢包特性吸取了FC网络的可靠性和以太网的灵活扩展能力。