**Infiniband网络结构分析**

**智能1702班 尹淑青 201708010601**

1. Infiniband简介

nfiniBand（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。

nfiniBand技术不是用于一般网络连接的，它的主要设计目的是针对服务器端的连接问题的。因此，InfiniBand技术将会被应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。

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。

1. Infiniband协议简介

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)、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。

1. 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。

1. 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。

1. 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。

1. 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. 分析解读Infiniband架构和服务能力

**InfiniBand的软件架构：**

InfiniBand软件栈的设计是为了简化应用部署。IP和TCP套接字应用程序可以利用InfiniBand性能，而无需对运行在以太网上的现有应用程序进行任何更改。这同样适用于SCSI、iSCSI和文件系统应用程序。位于低层InfiniBand适配器设备驱动程序和设备独立API(也称为verbs)之上的上层协议提供了行业标准接口，可以无缝部署现成的应用程序。

InfiniBand特定模块和协议：

内核代码逻辑上分为三层：HCA驱动程序、核心InfiniBand模块和上层协议。用户级访问模块实现了必要的机制，允许从用户模式应用程序访问InfiniBand硬件。核心InfiniBand模块包括InfiniBand设备的内核级中间层，中间层允许访问多个HCA NICs并提供一组公共共享服务，包括：

中间层主要功能

通信经理(CM) --CM提供了允许客户建立连接所需的服务。

SA客户端——SA(子网管理员)客户端提供了允许客户端与子网管理员通信的功能。SA包含建立连接所需的重要信息，如路径记录。

SMA-子网管理器代理响应子网管理包，允许子网管理器在每个主机上查询和配置设备。

PMA -性能管理代理响应允许检索硬件性能计数器的管理包。

MAD服务——管理数据报(MAD)服务提供一组接口，允许客户端访问特殊的InfiniBand队列对(QP)， 0和1。

GSI -通用服务接口(GSI)允许客户端在特殊QP1上发送和接收管理包。

队列对(QP)——重定向高层管理协议，通常将共享对特殊QP1的访问重定向到专用QP。这是为带宽密集型的高级管理协议所需要的。

SMI -子网管理接口(SMI)允许客户端在特殊QP0上发送和接收数据包。这通常由子网管理器使用。

Verbs-对中间层提供由HCA驱动程序提供的Verbs访问。InfiniBand体系结构规范定义了Vbers。Vbers是必须提供的函数的语义描述。中间层将这些语义描述转换为一组Linux内核应用程序编程接口(API)。

中间层还负责在异常程序终止或客户端关闭后，对没有释放的已分配资源的资源跟踪、引用计数和资源清理。

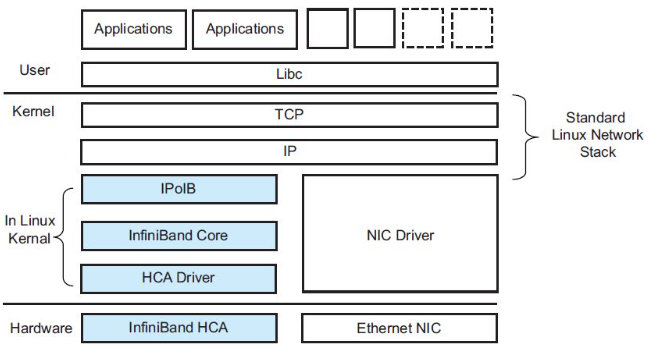
InfiniBand堆栈的最低层由HCA驱动程序组成。每个HCA设备都需要一个特定于HCA的驱动程序，该驱动程序注册在中间层，并提供InfiniBand Verbs。

如IPoIB，SRP，SDP，iSER等高级协议，采用标准数据网络，存储和文件系统应用在InfiniBand上操作。除了IPoIB提供了InfiniBand上TCP/IP数据流的简单封装外，其他更高级别的协议透明地支持更高的带宽、更低的延迟、更低的CPU利用率和端到端服务，使用经过现场验证的RDMA(远程DMA)和InfiniBand硬件的传输技术。下面将讨论这些高级协议，以及如何快速启用现有的应用程序对InfiniBand进行操作。

**IB对基于IP的应用支持**

在InfiniBand上评估任何基于IP的应用程序的最简单方法是使用上层协议IP over IB (IPoIB)。在高带宽的InfiniBand适配器上运行的IPoIB可以为任何基于ip的应用程序提供即时的性能提升。IPoIB支持在InfiniBand硬件上的(IP)隧道数据包。

如下图，在Linux中，协议是作为标准的Linux网络驱动程序实现的，这允许任何使用标准Linux网络服务的应用程序或内核驱动程序在不修改的情况下使用InfiniBand传输。Linux内核2.6.11及以上版本支持IPoIB协议，并对InfiniBand核心层和基于Mellanox技术公司HCA的HCA驱动程序的支持。



这种在InfiniBand上启用IP应用程序的方法对于带宽和延迟不重要的管理、配置、设置或控制平面相关数据是有效的。由于应用程序继续在标准TCP/IP网络栈上运行，应用程序完全不知道底层I/O硬件。然而，为了获得充分的性能并利用InfiniBand体系结构的一些高级特性，应用程序开发人员也可以使用套接字直接协议(SDP)和相关的基于套接字的API。

InfiniBand不仅对基于IP的应用提供了支持，同时对基于Socket、SCSI和iSCSI，以及对NFS的应用程序提供了支持。