《计算机系统设计》课程设计报告

湖南大學

HUNAN UNIVERSITY

选题名称：Infiniband 网络结构分析

姓　　名：王倩

学　　号：201708010630

专业班级：物联 1702

Infiniband 开放标准技术简化并加速了服务器之间的连接,同时支持服务器与远程存储和网络设备的连接。

什么是 InfiniBand 网络

InfiniBand 是一种网络通信协议，它提供了一种基于交换的架构，由处理器节点之间、处理器节点和输入/输出节点(如磁盘或存储)之间的点对点双向串行链路构成。每个链路都有一个连接到链路两端的设备，这样在每个链路两端控制传输(发送和接收)的特性就被很好地定义和控制了。

InfiniBand 通过交换机在节点之间直接创建一个私有的、受保护的通道，进行数据和消息的传输，无需 CPU 参与远程直接内存访问(RDMA)和发送/接收由 InfiniBand 适配器管理和执行的负载。

适配器通过 PCI Express 接口一端连接到 CPU，另一端通过 InfiniBand 网络端口连接到 InfiniBand 子网。与其他网络通信协议相比，这提供了明显的优势，包括更高的带宽、更低的延迟和增强的可伸缩性。

什么是 InfiniBand 架构

InfiniBand Architecture(IBA)是为硬件实现而设计的，而 TCP 则是为软件实现而设计的。因此，InfiniBand 是比 TCP 更轻的传输服务，因为它不需要重新排序数据包，因为较低的链路层提供有序的数据包交付。传输层只需要检查包序列并按顺序发送包。

进一步,因为 InfiniBand 提供以信用为基础的流控制(发送方节点不给接收方发送超出广播 “信用“大小的数据包),传输层不需要像 TCP 窗口算法那样的包机制确定最优飞行包的数量。这使得高效的产品能够以非常低的延迟和可忽略的 CPU 利用率向应用程序交付 56、100Gb/s 的数据速率。

IB 是以通道(Channel)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)，所以会有所谓的 IBA 交换器(Switch)，此外在线路不够长时可用 IBA 中继器(Repeater)进行延伸。

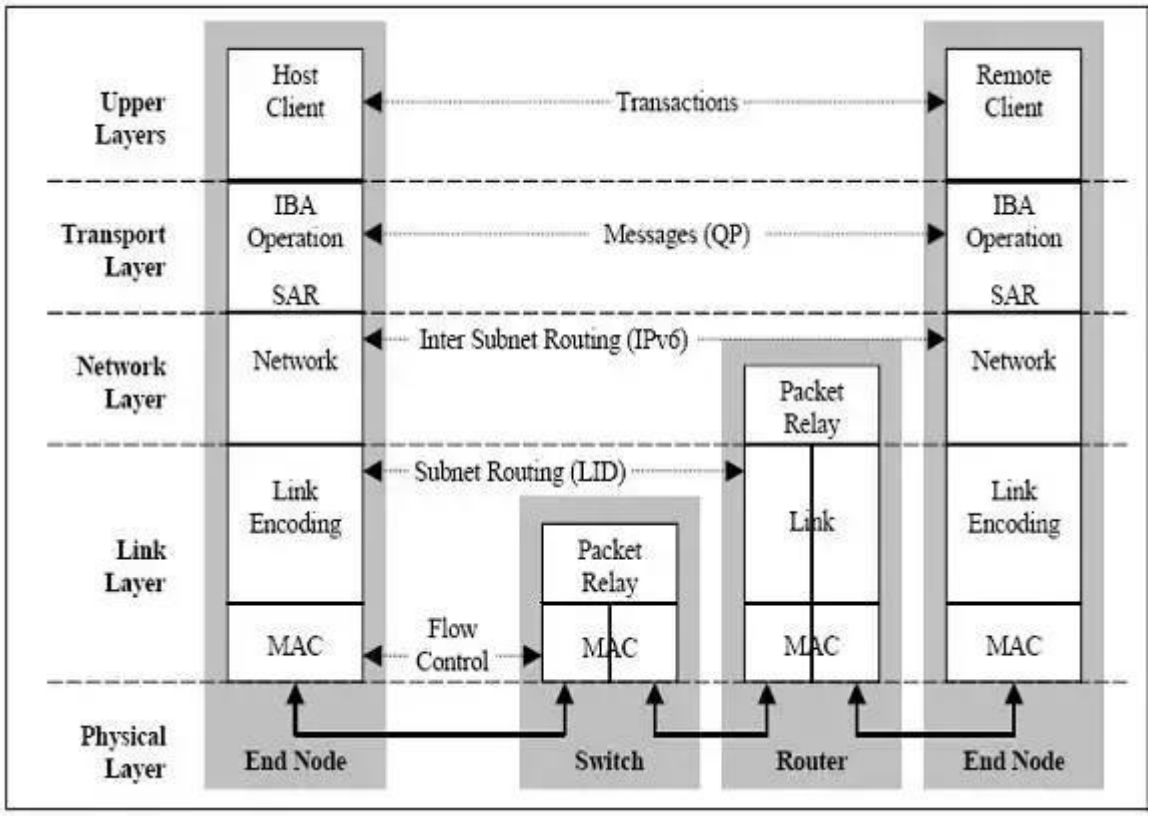
而每一个 IBA 网络称为子网(Subnet)，每个子网内最高可有 65,536 个节点(Node)，IBASwitch、IBA Repeater 仅适用于 Subnet 范畴，若要通跨多个 IBA Subnet 就需要用到 IBA 路由器(Router)或 IBA 网关器(Gateway)。

至于节点部分,Node 想与 IBA Subnet 接轨必须透过配接器(Adapter)，若是 CPU、内存部分要透过 HCA (Host Channel Adapter)，若为硬盘、I/O 部分则要透过

TCA (Target Channel Adapter)，之后各部分的衔接称为联机(Link)。上述种种构成了一个完整的 IBA。

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。

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。

- 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。

InfiniBand 技术的发展

1999 年开始起草规格及标准规范,2000 年正式发表,但发展速度不及 Rapid I/O、PCI-X、PCI-E 和 FC，加上 Ethernet 从 1Gbps 进展至 10Gbps。所以直到 2005 年之后，InfiniBand Architecture(IBA)才在集群式超级计算机上广泛应用。全球 Top 500 大效能的超级计算机中有相当多套系统都使用上 IBA。

随着越来越多的大厂商正在加入或者重返到它的阵营中来，包括 Cisco、IBM、HP、Sun、NEC、Intel、LSI 等。InfiniBand 已经成为目前主流的高性能计算机互连技术之一。为了满足 HPC、企业数据中心和云计算环境中的高 I/O 吞吐需求，新一代高速率 56Gbps 的 FDR (Fourteen Data Rate) 和 EDR InfiniBand 技术已经出现。

InfiniBand 技术的优势

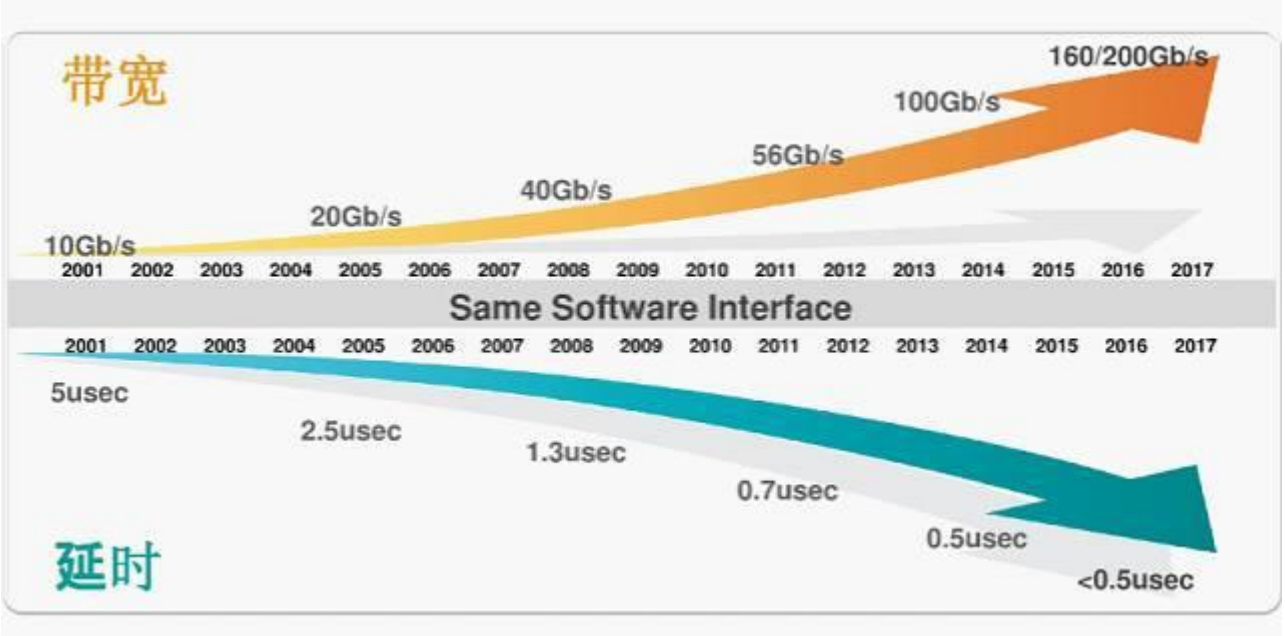
Infiniband 大量用于 FC/IP SAN、NAS 和服务器之间的连接,作为 iSCSI RDMA 的存储协议 iSER 已被 IETF 标准化。目前 EMC 全系产品已经切换到 Infiniband 组网，IBM/TMS 的 FlashSystem 系列，IBM 的存储系统 XIV Gen3，DDN 的 SFA 系列都采用 Infiniband 网络。

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。

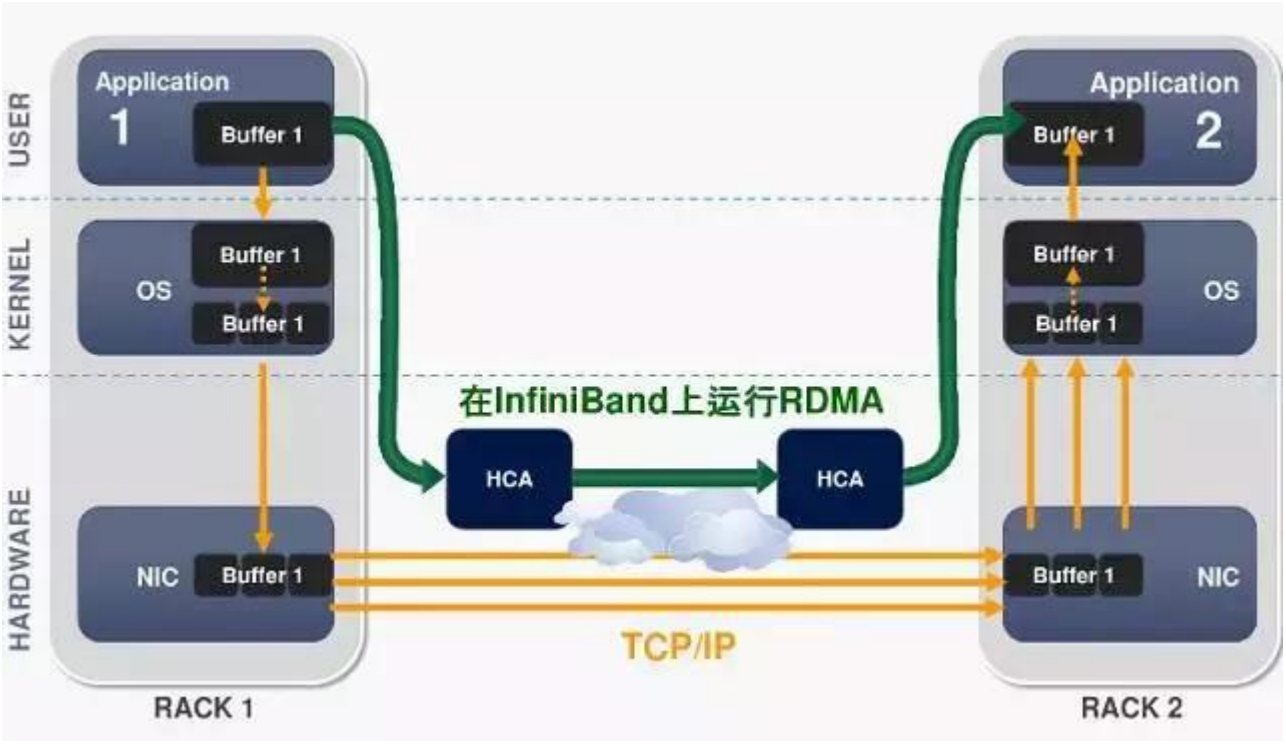
存储系统已不能满足于传统的FC SAN所提供的服务器与裸存储的网络连接架构。HP SFS和IBM GPFS 是在Infiniband fabric连接起来的服务器和iSER Infiniband存储构建的并行文件系统，完全突破系统的性能瓶颈。

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微妙、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)。



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。

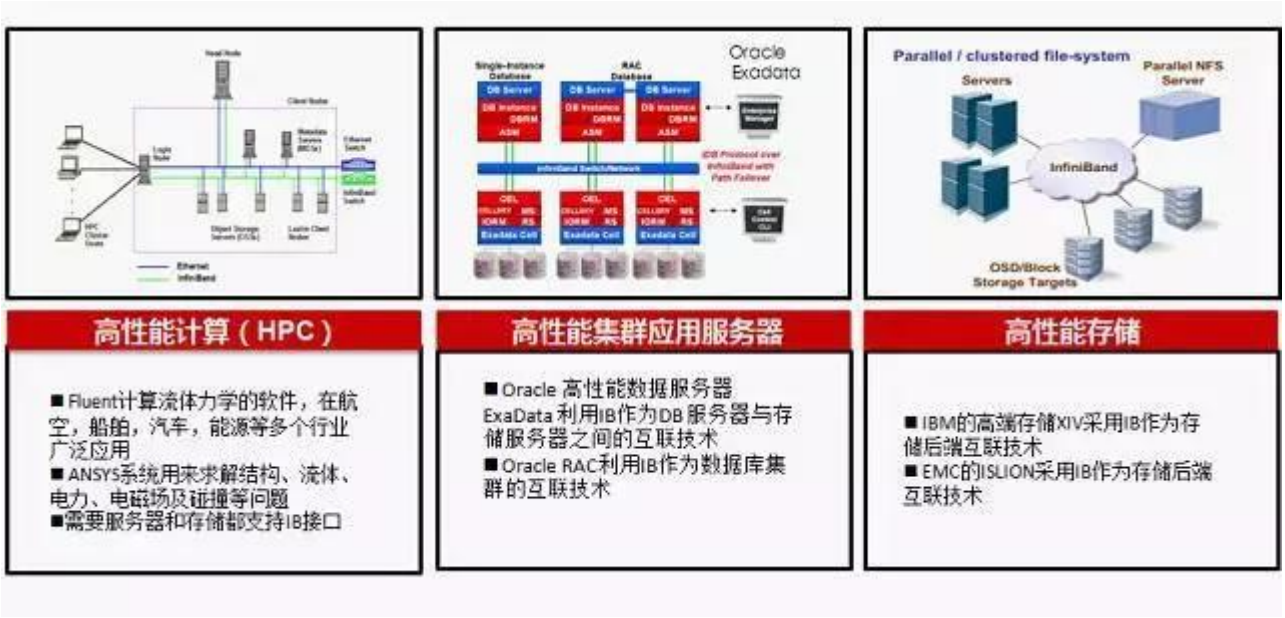
InfiniBand 速率发展介绍

InfiniBand 串行链路可以在不同的信令速率下运行，然后可以捆绑在一起实现更高的吞吐量。原始信令速率与编码方案耦合，产生有效的传输速率。编码将通过铜线或光纤发送的数据的错误率降至最低，但也增加了一些开销(例如，每 8 位数据传输 10 位)。

典型的实现是聚合四个链接单元(4X)。目前，InfiniBand 系统提供以下吞吐量速率:

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。

InfiniBand 管理软件

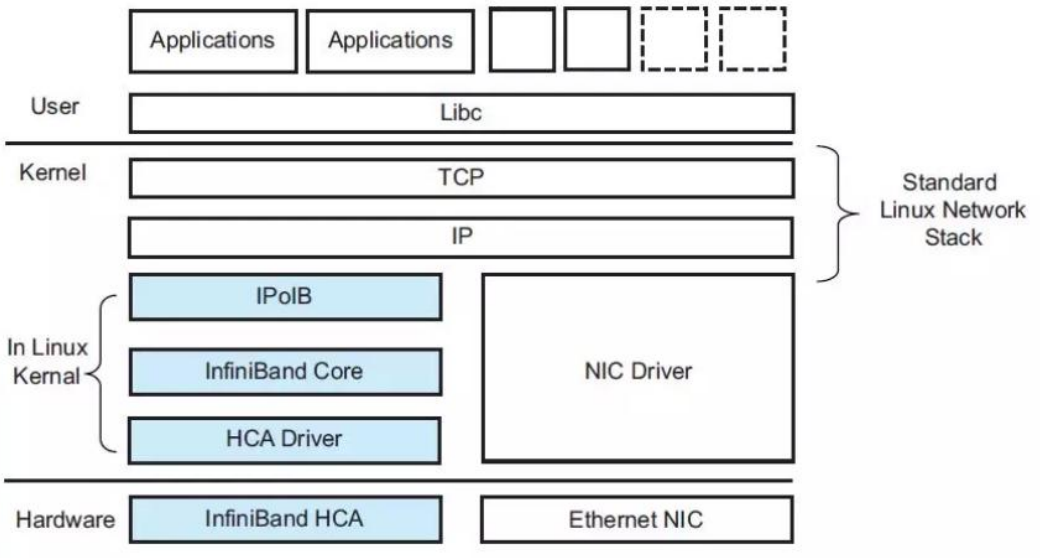
OpenSM 软件是符合 InfiniBand 的子网管理器(SM)，运行在 Mellanox OFED 软件堆栈进行 IB 网络管理，管理控制流走业务通道，属于带内管理方式。

OpenSM 包括子网管理器、背板管理器和性能管理器三个组件，绑定在交换机内部的必备部件。提供非常完备的管理和监控能力，如设备自动发现、设备管理、Fabric 可视化、智能分析、健康监测等等。

InfiniBand 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。

例如，在 iSER 协议中，采用了 SCSI 中间层的方法插入到 Linux，iSER 在额外的抽象层(CMA，Connection Manager Abstraction layer)上工作，实现对基于 InfiniBand 和 iWARP 的 RDMA 技术的透明操作。

这样使得采用 LibC 接口的用户应用程序和内核级采用 Linux 文件系统接口的应用程序的透明化，不会感知底层使用的是什么互连技术。