Infiniband 网络架构分析

1. InfiniBand 简介

InfiniBand 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。

InfiniBand 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。

2. InfiniBand 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。该软件由一组内核模块和协议组成。还有一些关联的用户模式共享库，这些库在图中没有显示。在用户级操作的应用程序对底层互连技术保持透明。

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。

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。

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。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。