Infiniband网络结构分析

智能1702 201708010528 莫诗幻

1. Infiniband简介

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。

1. 基本特征

1、Infiniband协议的主要特点是高带宽(现有产品的带宽4xDDR 20Gbps,12x DDR 60Gbps, 4xSDR 10Gbps, 12xSDR 30Gbps、预计两年后问世的QDR技术将会达到4xQDR 40Gbps,12x QDR 120Gbps)、低时延(交换机延时140ns、应用程序延时3μs、一年后的新的网卡技术将使应用程序延时降低到1μs水平)、系统扩展性好(可轻松实现完全无拥塞的数万端设备的Infiniband网络)。另外Infiniband标准支持RDMA(Remote Direct Memory Access)，使得在使用Infiniband构筑服务器、存储器网络时比万兆以太网以及Fibre Channel具有更高的性能、效率和灵活性。

2、该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。

2、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。

3、采用Intel架构的处理器的输入/输出性能会受到PCI或者PCI-X总线的限制。总线的吞吐能力是由总线时钟决定的（比如33.3MHz，66.6MHz 以及133.3MHz）和总线的宽度（比如32位或者64位）。在最通常的配置中，PCI总线速度被限制在500 MB /秒，而PCI-X总线速度被限制在1 GB/秒。这种速度上的限制制约了服务器和存储设备、网络节点以及其他服务器通讯的能力。在InfiniBand的技术构想中，InfiniBand直接集成到系统板内，并且直接和CPU以及内存子系统互动。但是，在短期内，InfiniBand支持将由PCI和PCI-X适配器完成；这样，InfiniBand在最初将会受到总线的制约。在2002年年底，InfiniBand技术将会完全被整合在Intel服务器供应商以及Sun生产的服务器中（80%的可能性）

　4、InfiniBand的扩展性非常高，在一个子网内可支持上万个节点，而每个网络中可有几千个子网，每个安装的系统中可以有多个网络结构。InfiniBand交换机通过子网路由分组，InfiniBand路由器将多个子网连接在一起。相对以太网，InfiniBand可以更加分散地进行管理，每个子网内有一个管理器，其在路由分组、映射网络拓扑、在网络内提供多个链路、监视性能方面起决定性的作用。子网管理器也能保证在特别通道内的带宽，并为不同优先权的数据流提供不同级别的服务。子网并不一定是一个单独的设备，它可以是内置于交换机的智能部件。

1. InfiniBand构架：

1、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。

1. 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。
2. InfiniBand网络主要上层协议
3. SDP (Sockets Direct 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。
4. 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。
5. iSER(iSCSI RDMA 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。
6. RDS(Reliable Datagram 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。
7. 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。
8. uDAPL(User Direct 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。
9. iSER (iSCSI Extensions for RDMA)和NFSoRDMA (NFS over RDMA)，SRP (SCSI RDMA Protocol) 等是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。