湖南大學

HUNAN UNIVERSITY

计算机设计

学生姓名	周珍冉
学生学号	201708010610
专业班级	智能 1702
指导老师	吴强
完成日期	2019.12.10

Infiniband 网络结构分析

一、 Infiniband

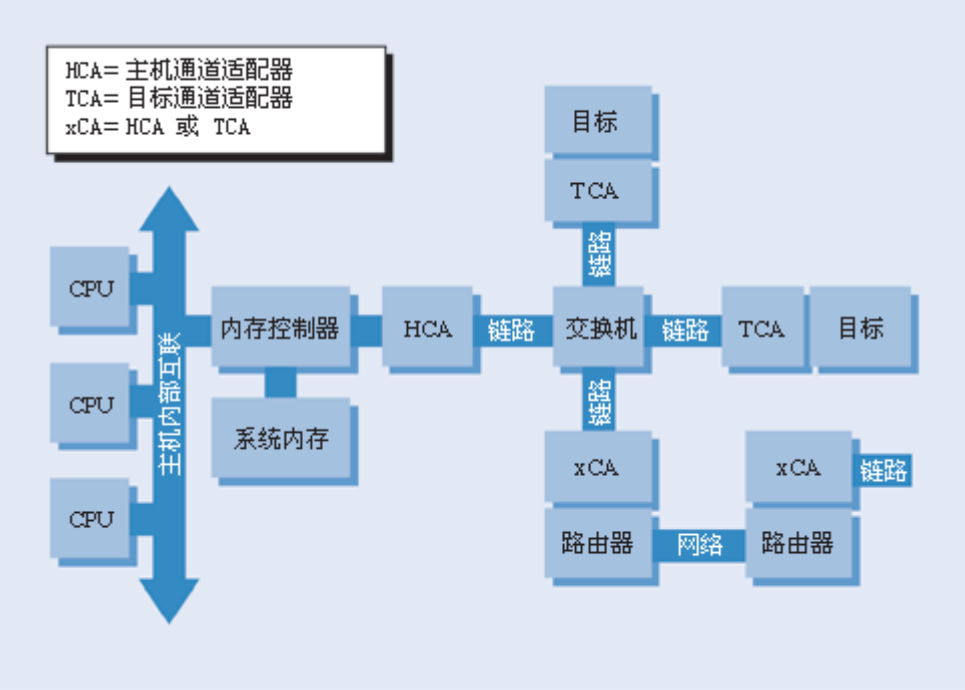
InfiniBand（IB）是一个用于高性能计算的计算机网络通信标准，它具有极高的吞吐量和极低的延迟，用于计算机与计算机之间的数据互连。InfiniBand 技术的主要设计目的是针对服务器端的连接问题的。因此，InfiniBand 技术将会被应用于服务器与服务器（比如复制，分布式工作等），服务器和存储设备（比如 SAN 和直接存储附件）以及服务器和网络之间（比如 LAN， WANs 和 the Internet）的通信。

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。

二、 InfiniBand 网络结构

IB 网络互连结构:



InfiniBand 网的组成元素主要有:

✧ HCA(Host Channel Adapter)（主机通道适配器）

主机结点上有一个(或多个)HCA, 它是驻留在处理节点上的一个网络接口, 它通过一条或几条双向链路连接内存控制器到交换结构, 主机通道适配器包含一个链路协议引擎, 用它来执行硬件链路中的协议。其作用是在 IB 网络和主机 CPU 之间建立数据传输通道。

✧ TCA(Target Channel Adapter)（通道适配器）

TCA 提供到 I / O 控制器的链路和传输服务，使 I / O 设备可脱离主机而直接置于网络中。TCA 用作一个外部 RAID 子系统的前端接口，它还可以被用作一种接口，与其他形式的 I/O（如并行 SCSI、以太网或光纤通道）连接。通道适配器是可以在不中断 CPU 运行的情况下处理所有 I/O 功能的智能设备，具有自动路

径迁移功能。我们将 HCA 和 TCA 统称为 CA。

✧ Switch（交换器）

Switch 是组成 IBA 网络的关键部件，它允许多个 CA 和它连接，并且处理 CA 之间的通信。它具有分区功能及自动路径迁移功能，用以组建子网，并实现子网内路由。

✧ Router（路由器）

路由器用于连接不同的子网。CA 所驻留的节点、Switch、Router 统称为网络节点。InfiniBand 结构通过交换机和路由器连接主机和设备，每个主机通道适配器、目标通道适配器、交换机、和路由器都有一个全球唯一的基于 Ipv6 的标识符。

✧ Infiniband 网络管理

网络管理对子网进行配置并维护其正常运行，同时向上层程序提供服务。IBA 的网络管理分成两个部分来说明：子网管理(Subnet Management)和通用服务管理(General Service Management)。子网管理主要负责初始化和配置子网，并且维护子网的正常运行，它通过子网内专门的网络管理程序 SM(SubnetManager)施行。通用服务管理提供高层程序所需的服务，它包括：SA(Subnet Administrator)、PM(Perfor . lilance Mmmgenmnt)、DM(Device Management)、BM(Basebord Management)等。其中，SM 和 SA 是网络管理的重要部件，由于两者关系密切，所以在具体实现上将它们绑定在一起。

三、 Infiniband 和分布式系统的区别

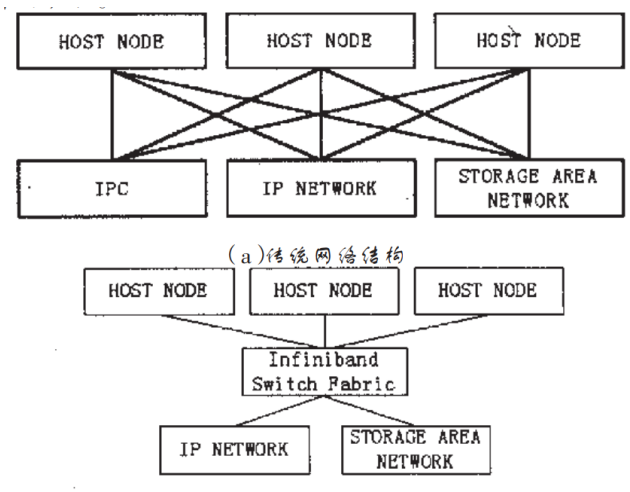
当今分布式系统的的一个特点就是有多套功能特定的网络结构：用作进程间

通信的网络、IP 网络、存储域网络，每种结构都有自己的发展趋势、管理策略和接口。这些不同的网络技术通过不同的网络接口连接到每个服务器上，最后连接到服务器内的共享输入输出总线结构上，这导致了 I/O 瓶颈和不同资源之间额外的通信开销。

一个独立的 I/O 体系应该简化现有的互连方式并提供一个灵活的框架以适应计算和存储技术的发展。迎合不断增长的用户需求、并允许软硬件制造商提供新型的兼容性的解决方案的唯一方法是开发一种通用的 I/O 体系结构。因此，IBTA 的目标是结合现有互联体系结构的优点形成一个标准，即 Infiniband Archic。

该标准提供了一个统一的结构用来连接和共享多个模块设备上的资源，从服务器集群到存储域网到用户网络，这将简化主机和资源之间的连接并允许它们分别进入功能特定的子网。

下图为现今传统网络结构和 Infiniband 结构的比较，对 a 图说进程向通信（Interprocessing communication）、IP 网络和存储域网分别拥有不同的网络，而在 b 图 InfiniBand 结构中，只有一种网络通过 InfiniBand switch 连接却达到了与 a 图同样的功能。



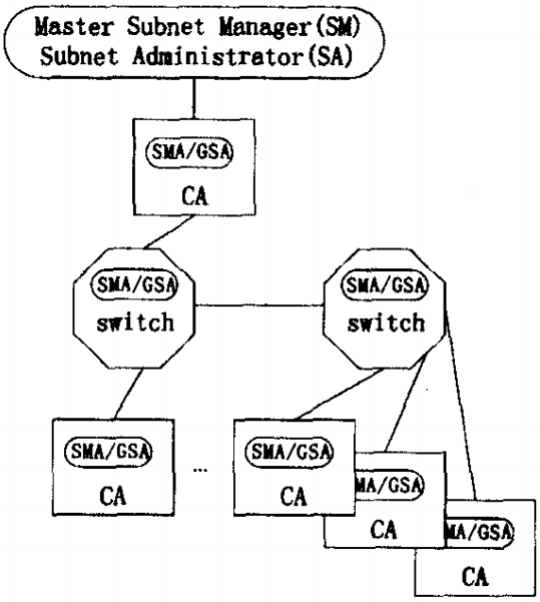
（a）传统网络结构

（b）InfiniBand 结构

图 2 传统网络结构和 InfiniBand 结构的比较

四、 Infiniband Archic

IBA 的网络管理模型如下图所示。



每个子网可以同时存在多个 SM，其中的一个为 Master SM(主 SM)，其他的为 Standby SM(备用 SM)。只有 MasterSM 才能对子网进行配置和管理。每个 Switch、CA 和 Router 上都驻留一个子网管理代理 SMA(SM Agent)和通用服务管理代理(GS Agent)。SMA 由 Master SM 管理，支持 SM 对各个网络节点信息的收集和配置；SA 响应 GSA 发来的服务请求并作相应处理。IBA 规范专门为网络管理定义了网络管理包 MAD(Management Datagram)。MAD 分成两种：用于子网管理的管理包 SMP(SM Packet)和用于通用服务管理的管理包 GMP(GS Management Packet)。

网络管理程序运行在任何网络管理节点上，网络管理程序必须和所运行的网络节点的某个端口绑定。网络一启动，每个端口都处于 DOWN 状态，只能发送和接收管理消息包，不能正常地收发数据包。此时网络内的 SM 使用定向路由包探测子网。当网络内有多个 SM 存在时，在初始化时竞选产生一个 Master SM，

其余的是备用 SM(Standby SM)。只有 Master SM 才能配置子网。具体实现中，如果某个 SM 成为 Master SM，和该 SM 绑定在一起的 SA 开始响应各个节点来的服务请求。

在网络未配置好之前，管理包的收发使用定向路由。在配置 IBA 网络时需要给 CA、Router 的每个端口和 Switch 分配 LID 号，正常的消息包收发时，通过消息包中的 LID 号识别源和目的；分配好 LID 号以后，为了进行消息包路由，需要建立所有 LID 之间的路由信息，并且将这些路由信息设置到 Switch 上；设置好路由信息以后要激活子网内所有端口，以进行正常的数据收发。所以，Master SM 主要负责以下工作：①查找子网的物理拓扑结构；②为每个 CA 的端口、Router 的端口和 Switch 分配 LID；③建立 LID 间的路由信息并且把路由表设置到 Switch 上；④激活端口；⑤定期扫描子网，探测子网结构的变化，实现对子网的动态维护。

Master SM 做完上述工作以后，建立一个包括所有子网信息的数据库 SMDB(SM Database)。各个节点如果想查询整个子网的信息，通过发送 SA 类型的 MAD 到 SA。SA 查询 SMDB，将需要的信息再通过 SA 类型的 MAD 发送到各个节点。

SA 实现以下功能：查询子网信息、事件跟踪、服务登记与撤销、多播操作

五、 总结心得

通过查询资料，我了解了 InfiniBand 的主要组成为 HCA、TCA、Switch、Router，其特点是高带宽、低时延、系统扩展性能好。Infiniband 标准支持 RDMA，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。了解学习了很多新知识、概念以及结构思想，收获很大。