章节

InfiniBand、iWARP 和 RoCE

11

马诺杰·瓦德卡

QLogic 公司研究员、首席技术专家

11.1简介

InfiniBand (IB) 是一种点对点互连。其零复制和远程直接内存访问 (RDMA) 等功能可直接将数据从发送方内存传输到接收方内存，而无需涉及主机处理器，从而有助于减少处理器开销。本章涵盖整体 IB 架构 (IBA) 及其各个层。本章的重点是链路层和网络层。随着 IB 不断发展以通过以太网为低延迟应用程序提供连接，互联网广域 RDMA 协议 (iWARP) 和 RoCE 正在成为向应用程序提供 RDMA 功能的有吸引力的选择。

本章详细介绍了这两个协议。

11.2 InfiniBand架构

IBA 定义了一种交换通信结构，允许许多设备在受保护的远程管理环境中同时以高带宽和低延迟进行通信。终端节点可以通过多个 IBA 端口进行通信，并且可以利用通过 IBA 结构的多条路径。

图 11.1演示了 IB [10]系统网络中的各种组件。
该网络由通过级联交换机和路由器连接的各种处理器节点和 I/O 单元组成。它允许处理器间通信的低延迟互连，支持存储设备到存储设备的连接，并且还演示了路由器可用于将连接扩展到广域网 (WAN)、局域网（LAN，通过以太网）或存储区域网络 (SAN)。路由器还提供多个 IB 子网之间的连接。

IBA 还定义了允许与小型计算机系统接口 (SCSI)、以太网和光纤通道 (FC) 等其他第 2 层技术进行通信的架构组件。

IBA 定义了一个分层协议，指定物理层、链路层、网络层、传输层和上层。它定义了通过各种介质进行的通信，包括印刷电路板 (PCB)、铜缆和光缆。 IB 允许三种链接速度：

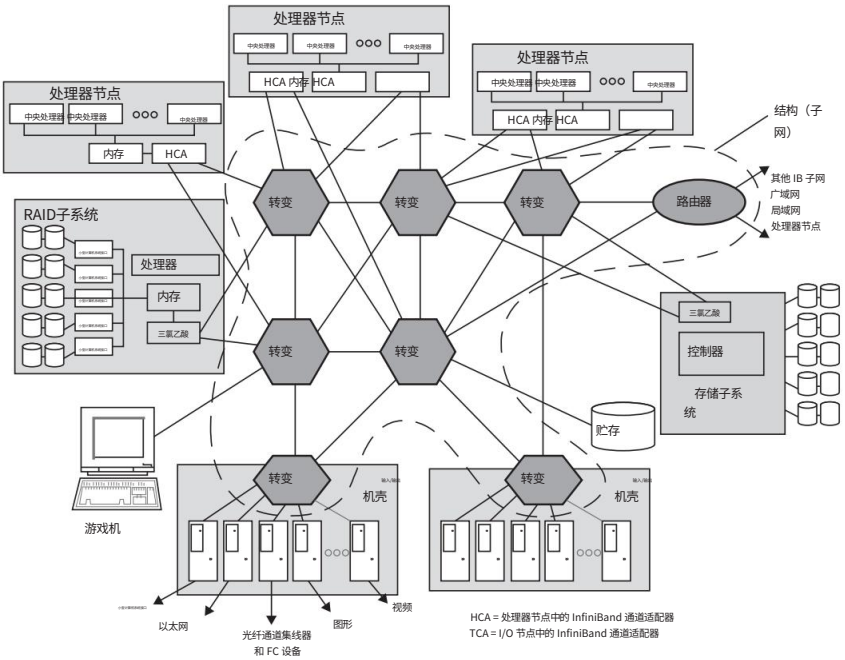


图11.1

IB 织物和组件。

超过 4 线（1X:单通道）、16 线（4X:4 通道)或 48 线（12X:12 通道）。
因此,如果单通道（4 根线:差分对 RX1 /RX2 和 TX1 /TX2)的速率为 2.5 Gbps,则 4X 连接提供 10 Gbps,12X 连接提供 30 Gbps 的通信。

IBA 支持节点之间的单播和多播流量。这种业务可以以可靠或不可靠的模式进行。它还支持连接或数据报模式进行通信。架构内提供了各种QoS（服务质量)机制,以保证网络中的无损和差异化流量。以下部分将更详细地描述 IBA 的所有这些方面。

11.3 IB 网络11.3.1

网络拓扑IBA 网络由多个可

以通过路由器连接的子网组成。每个端节点可以是处理节点、I/O 单元或存储子系统。 IBA 允许使用 RDMA 协议在这些参与节点之间进行通信。这可实现极低延迟的数据传输以及低 CPU

进程间通信 (IPC) 应用程序的利用。远程数据放置直接在源节点和目标节点之间实现,因此避免了数据复制并最大限度地减少了操作系统的参与。这些因素共同减少了总体延迟以及 CPU 开销（图 11.2）。

任何IB设备都可以与一个或多个IB设备或交换机连接。
一条或多条链路可用于此类连接。

11.3.2子网组件11.3.2.1通道适配器

IBA 定义了驻留在服务器或 I/O 系
统中的两种类型的适配器。如果适配器驻留在主机系统（例如服务器)中,则称为主机通道适配器（HCA）。如果适配器驻留在存储目标系统中,则称为目标通道适配器。

通道适配器为操作系统和应用程序提供与物理端口的连接。 HCA 向操作系统提供接口,并提供 IB 定义的所有动词接口。 Verb 是应用程序和通道适配器提供的功能之间的抽象接口。

每个适配器可以有一个或多个端口。每个端口通过“虚拟通道”(VL) 进一步区分流量。每个VL 都可以独立进行流量控制。如图 11.3所示， DMA 可以从本地和远程应用程序启动。

通道适配器携带一个称为全局唯一标识符 (GUID) 的唯一地址。适配器的每个端口还分配有一个唯一标识符,即端口 GUID。 GUID 由适配器供应商分配。给定子网的管理实体（称为子网管理器 (SM))将本地标识符 (LID) 分配给通道适配器的每个端口。

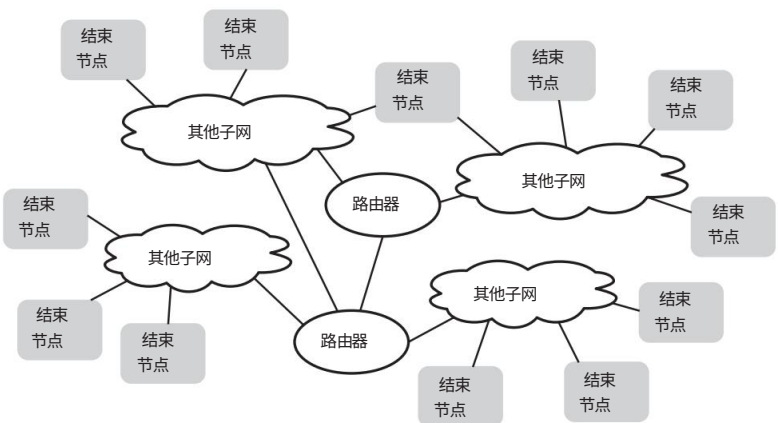


图11.2

其他网络组件。

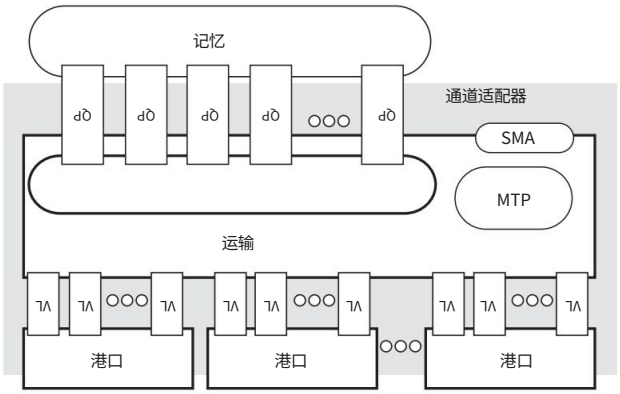


图11.3

IB 通道适配器。

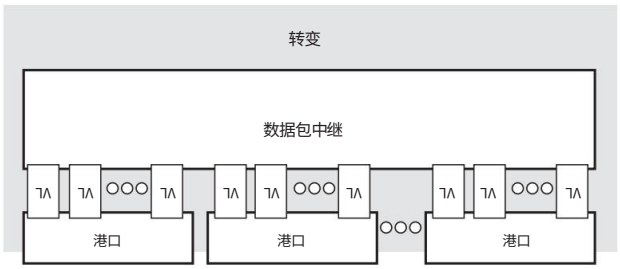


图11.4

IB 开关。

11.3.2.2交换机交换机

包含多个IB 端口。它根据每个数据包的本地路由标头 (LRH) 中的目标地址在适配器端口之间转发数据包。

交换机将单播数据包转发到单个目标端口，并将多播数据包转发到其转发数据库中配置的多个端口。交换机使用的转发数据库由SM配置。交换机只转发数据包，不修改数据包的标头（图 11.4）。

11.3.2.3路由器与交换

机类似，路由器将数据包从源转发到目的地。然而，路由器根据全局路由头（GRH）转发数据包。它们在从一个子网转发到另一子网时修改 LRH。每个子网都用子网前缀进行标识。路由器使用 IPv6 指定的协议交换路由信息（图 11.5）。

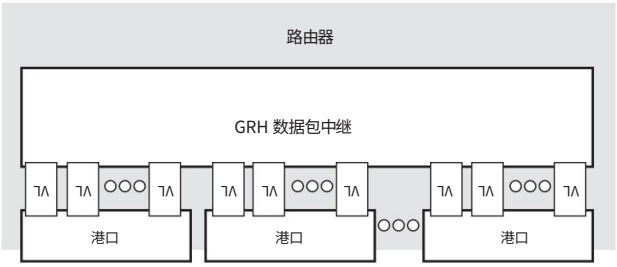


图11.5

IB路由器。

源节点指定路由器的 LID 和数据包发送到的目的地的全局标识符 (GID)。每个路由器使用子网信息和路由表将数据包转发到下一个路由器。执行路由直至数据包到达目标子网。最后一个路由器使用与目标 GID 关联的本地 ID 将数据包转发到最终目标。

11.3.2.4子网管理器SM 配置
子网中的本地组件。它向子网中的所有节点提供LID,并且还向子网中的交换机提供转发信息。 SM 通过子网管理代理 (SMA) 与子网内的所有节点进行通信。每个IB节点都需要实现SMA。

子网中可以有多个 SM,但在给定时间只能有一个处于活动状态。所有非活动/备份 SM 都维护活动 SM 的转发信息的副本,并在活动 SM 出现故障时使用它继续向子网提供管理服务。

11.4通信机制11.4.1通信服务（传输）

IBA根据应用的需要支持不同类型的IB节点之间的通信机制。

11.4.1.1可靠连接和可靠数据报为了实现可靠通信,数据通过序列
号和确认消息（ACK/NAK)的组合来可靠地传送。在检测到数据包错误或丢失时,源可以通过重新传输数据包来恢复,而无需用户应用程序的参与。这种模式保证消息包只传送一次。当应用程序需要依赖底层传输来保证消息传送到目的地时,可以使用此模式。

此模式释放应用程序以防范底层媒体或传输机制的不可靠性。

可靠连接 (RC) 模式使用源端节点和目标端节点之间的直接专用连接来提供节点之间的可靠数据传输。

可靠数据报 (RD) 模式向任何端节点提供可靠的数据包消息传送，而无需源端节点和目标端节点之间的专用连接。这是一个可选模式。

11.4.1.2不可靠数据报和不可靠连接不可靠模式对于对数据包丢失不敏感或能够自行处理数据包丢失但需要快速数据传输的应用程序非常有用。在这种模式下，不保证数据从源节点到目的端节点的传输。

在不可靠数据报（UD)模式下，数据可以从源节点发送到目的端节点，而无需建立任何连接。不保证数据包送达。在此模式下，不会检测到数据丢失。

不可靠连接（UC)模式是在源端节点和目的端节点之间建立专用连接，在没有传输保证的情况下进行消息传递。错误（包括序列错误)被检测并记录，并且不会通知回源端节点。

11.4.1.3原始IPv6 数据报和原始以太网类型数据报这是UD 的特殊模式，其中仅使用本地传输头信息。非 IBA 传输层使用此模式在 IB 网络上建立隧道数据。

11.4.2节点间的寻址

IB 通信需要对每个可寻址实体（节点、卡、端口、端口内的队列对（QP)等)进行唯一标识，以便可以正确传送数据包。此类数据包可能是子网内通信的一部分，也可能属于通过路由器跨子网的流。

当通信恰好在两个可寻址实体之间时，流可以是单播的。多播流用于多个实体之间的通信（图 11.6）。

LID：LID是16位单播或多播标识符，在子网内是唯一的；它不能用于子网之间的路由。 LID由SM分配。 LID 包含在每个数据包的 LRH 内。

GID：GID 是一个 128 位单播或多播地址，并且在全球范围内是唯一的，这使得它可以用于跨子网路由数据包。 GID 是有效的 IPv6 地址，具有 IBA 定义的附加限制。

GID分配范围从默认分配（根据制造商分配的标识符计算)到SM分配的地址。

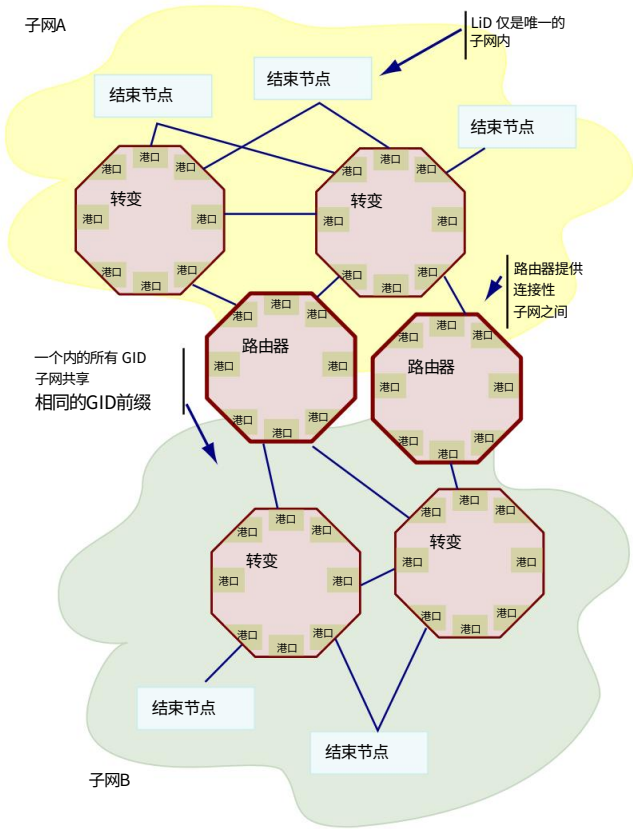


图11.6

IB 寻址范围。

单播标识符:单播 LID 或 GID 标识单个可寻址实体。 A
寻址到单播标识符的数据包将被传送到单个端端口。
组播标识符:组播 LID 或 GID 标识一组可寻址的组播地址。
终端端口。发送到多播标识符的数据包必须传递到所有
作为该标识符一部分的结束端口。

11.4.3数据包格式

IB 网络中定义了两类数据包。

IBA数据包:携带传输标头的IB数据包在IBA结构上路由
并使用 IBA 本地交通设施。
原始数据包:这些数据包通常用于传输非 IBA
IB 网络上的数据包。所以这些数据包不包含IBA传输
标头。

图11.7

IB 数据包格式。

IBA 定义的数据包格式如图 11.7 所示。

所有数据包都需要本地路由标头（8 字节）。该标头用于在本地子网内转发数据包[11]。

需要路由到不同子网的所有数据包以及所有多播数据包都需要 GRH（40 字节）。 LRH 中的链路下一个标头字段指示数据包中是否存在 GRH 标头。

原始数据包仅包含 LRH 和原始或 IPv6 路由标头。

IBA 支持数据包中的两种类型的 CRC（循环冗余校验）。不变 CRC（4 字节）涵盖当数据包通过结构从源节点移动到目标端节点时不会更改的所有字段。变体 CRC (VCRC)（2 字节）涵盖数据包中的所有字段。每个 IBA 数据包都携带不变的 CRC，后跟 VCRC。原始数据包仅携带 VCRC。

11.5分层架构分层架构允许使用不同的组件构建解决方案，这些组件可以灵活地实现互操作，同时由于层之间定义良好的接口而具有操作的正确性。它还为系统中的不同功能块提供架构清晰度和分离，这有助于应用程序通过各种协议、网络和物理连接选项相互通信。这提供了多种可能的应用程序部署，而不需要对实现进行自上而下的更改。人们可以通过铜缆或光缆运行系统，并更改物理介质，而无需更改物理层之上的任何协议。

类似地，人们可以在一个子网内或跨多个子网运行应用程序，而无需应用程序意识到通信系统之间的网络隔离（图 11.8）。

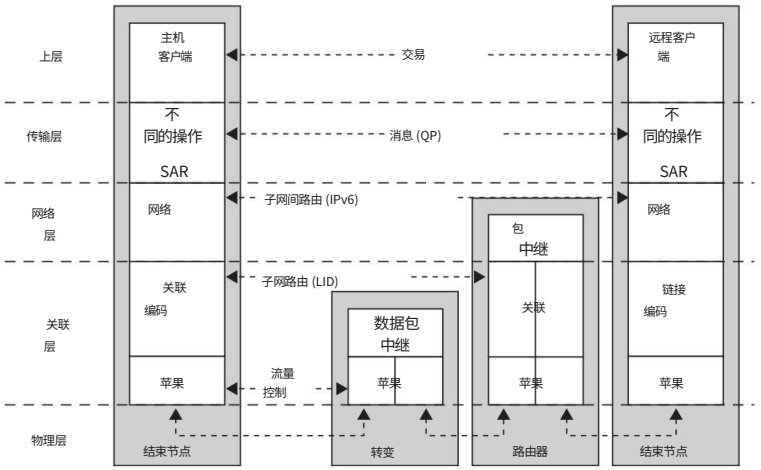


图11.8

IB分层架构。

IBA 建立了一个跨五层的分层架构。顶层提供应用程序接口,而底层定义系统之间的物理连接。

11.5.1物理层物理层定

义实际比特如何在系统之间的物理连接上流动。系统可以通过背板或光纤或铜介质相互连接。电缆的长度可以变化,连接到系统的机械连接器也可以变化。 IB 的物理层定义了此类连接的电气和机械方面。

物理层负责从链路层接收（控制和数据)字节,以电或光形式将它们发送到链路对等体,然后将接收到的字节传送到接收器处的链路层。它为链路层提供了有关连接到链路对等点的实际物理介质的透明度。

物理层还负责与链路对等点进行底层物理介质的速度和宽度协商。

IBA 定义了四种类型的物理连接选项来连接 IB 设备。
它提供了所有这些的电气、光学和机械规格。下面列出了物理连接选项：

1. 背板端口 2. 铜口
3. 光纤端口 4.
主用线缆端口

11.5.1.1数据包格式和链路宽度数据包由线路上

称为 SDL（数据包分隔符开始)和 EGP（好数据包分隔符结束)或 EBP（坏数据包分隔符结束)的特殊符号分隔。链路由两个节点之间的多个连接“通道”形成。具有单通道的端口称为 1X 端口。同样,具有 4 通道或 12 通道的端口分别称为 4X 或 12X。

对于具有多个通道的端口,数据包在所有通道上进行“字节条带化” （图 11.9）。

11.5.1.2速度和宽度协商IBA 以不同的速度运

行链路,因为新一代已经发展到以更高的速度运行链路。在 SDR（单一数据速率)中,信令速率为 2.5 Gbps。在接下来的几代中,速度已增加到 5 Gbps (DDR)、10 Gbps (QDR) 等。图 11.10显示了 IB 链路层的连续速度（也预测了未来的速度） [12]。

11.5.2链路层IB 架构中

的链路层定义了通过物理连接发送和接收数据包的机制。这些机制包括寻址、缓冲、流量控制、QoS、错误检测和交换。第 11.4.2 节更详细地讨论了寻址。

11.5.2.1数据包转发链路层定义

IBA 子网内数据包的转发。

在子网内,数据包使用 LID 转发。图 11.11显示了 LRH 的格式。这些标识符由 SM 在设备上配置,如第 11.3.2.4 节所述。交换机使用目标 LID 来查找给定数据包需要转发到的目标端口。

IBA 要求在流中的单播数据包（子网内相同源 LID 和目标 LID 之间的数据包)内维护按顺序数据包传送。

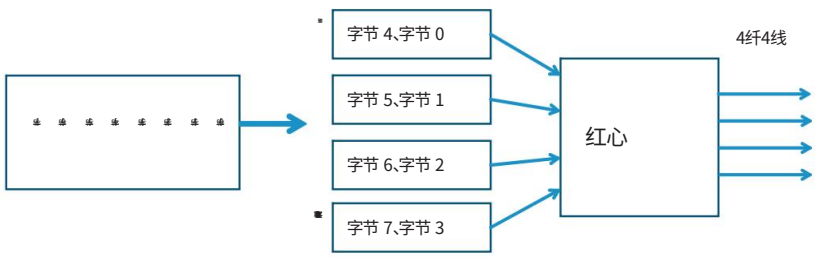
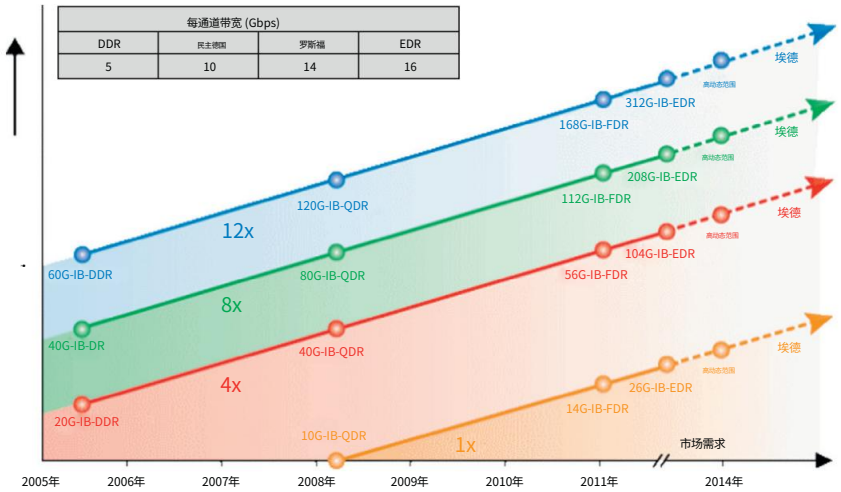


图11.9

跨通道对数据进行字节条带化。



SDR – 单一数据速率
DDR – 双倍数据速率
QDR 四倍数据速率
FDR – 14 数据速率
EDR – 增强的数据速率
HDR – 高数据速率
NDR – 下一个数据速率

图11.10

IBTA 信令速率路线图。

位 字节	31-24		23-16		15-8	7-0
0—3	VL	左室	SL	Rsv2 LNH	目的地本地标识符	
4—7	预留 5		数据包长度（11 位）		源本地标识符	

图11.11

本地路由标头 (LRH)。

这同样适用于多播和广播数据包。然而，有序要求不适用于节点接收的单播和多播数据包。

11.5.2.2数据完整性

IBA 使用 CRC 保证数据流经网络时的完整性
场地。

32 位不变 CRC 覆盖数据包中所有不变的字段
从头到尾。该字段不需要在每一跳都重新计算。这个领域
不存在于原始数据包中。

网络中变化的字段由 16 位 VCRC 覆盖。 VCRC 是
在每一跳进行计算，以捕获可能造成的任何数据完整性损害
导致在路由期间修改标头。该字段存在于所有数据包中，
包括原始数据包。

11.5.2.3虚拟通道IB 允许通

过称为 VL 的抽象在电线的物理连接上创建虚拟链路。就链路级流量控制等机制而言，每个 VL 可以彼此独立运行。这允许在给定物理端口或物理链路上区分不同“流”的流量。

可以将流分组为属于特定的“VL”，并且可以跨多个设备应用一致的流量工程规则，以为给定的 VL 提供特定的 QoS。

每个数据包在链路标头的 4 位 VL 字段中携带有关其属于哪个 VL 的信息。

每个VL都提供独立的缓冲资源，保证它们不干扰彼此的操作。每个 VL 的链路级流量控制可确保一个 VL 中的流量不会受到另一 VL 上断言的流量控制的影响。

IBA 在每个链路上定义了 15 个 VL。与 VL0 相比，VL15 被认为优先级更高。与较低优先级的流量相比，可以更迅速地为较高优先级的流量提供服务。可以根据流的 VL 规则将带宽/资源分配给流。管理流量使用最高优先级 VL，VL15。每个器件都需要支持 VL0 和 VL15。

11.5.2.4服务级别除了VL 之

外，IBA 还定义了一种为流分配QoS 标识符的机制 服务级别（SL）。这个 4 位字段包含在 LRH 中，它标识子网内给定流的 SL。当数据包穿过网络时，数据包中的该字段不会被修改。 SL 的实际含义有待实现，但是，它旨在由产品使用，以根据 SM 的规定为流提供流量差异化。

IBA 定义了将 SL 字段映射到给定端口的 VL 的机制。这种 SL 到 VL 的映射是通过映射表实现的，它允许 IB 节点根据映射表定义的规则和 VL 分配的固有期望为流提供 QoS。例如，映射到较高优先级 VL 的 SL 在给定端口获得较高优先级。

由于每个 SL 在给定端口上获得不同的调度，因此不同 SL 之间的顺序不保持不变。

11.5.2.5缓冲和流量控制IBA 提供了一种机制，

通过基于缓冲区到缓冲区信用的流量控制机制来保证数据包在链路上的无损传输。这要求每个接收器向发送器提供有关缓冲区可用性的信息，以便发送器可以将数据包传送到线路上的接收器。

由于有一个保留的缓冲区等待到达的数据包，因此接收方不会因为拥塞而丢弃数据包。

每个 VL 都需要在给定端口上具有单独的缓冲。这允许在每个 VL 上使用单独的流量控制。

IBA 定义了一种机制,让接收器通知发送器在给定时间点允许传输的数据量。 IBA 还指定协议以确保交换有关流量控制本身的信息的通信协议没有错误,并且可以在出现错误情况下从错误情况中恢复。

发送器和接收器定期重新同步其信息,以纠正给定 VL 上信用可用性信息中的任何不一致。

除了链路级流量控制之外,IBA 还指定了拥塞控制机制,允许网络中的拥塞端口请求流量的实际源放慢速度(与仅在链路级中传输前一个节点的端口相比)流量控制)。 IBA 指定了一种机制,其中可以检测 VL 上的拥塞,并由交换机针对违规数据包标记转发拥塞通知(FECN:转发显式拥塞通知)。该位由目的地解释并转为特殊管理数据包,用于向后通知(BECN:向后显式拥塞通知)到违规流的源。然后,源解释该数据包并暂时降低到给定拥塞目的地的数据注入速率(原始注入速率随着时间的推移而恢复)。

11.5.3网络层网络层提

供跨多个子网的路由。它指定跨 IBA 子网转发单播和组播数据包。这种路由可以由符合IBA以及非IBA(例如IP)规范的路由器来完成。

图 11.12中的 GRH 中提供的字段可用于此类路由。
通常,这些字段包括 SGID、DGID、TClass 和流标签(这些字段可以轻松映射到 IPv6 词汇表中,这是有意为之)。源全局标识符 (SGID) 和目标全局标识符 (DGID) 是可映射到 IPv6 地址的 128 位字段。路由的工作方式与 IP 路由非常相似,其中源子网内的目标本地标识符 (DLID) 将是

位 字节	31-24	23-16	15-8	7-0
0—3	IP版本	T级	流量标签	
4—7	支付宝		NxtHdr	霍普林特
8—11	SGID[127—96]			
12—15	滑动[95—64]			
16—19	SIGID[63—32]			
20—23	SIGID[31—0]			
24—27	DGID[127—96]			
28—31	DGID[95—64]			
32—35	DGID[63—32]			
36—39	DGID[31—0]			

图11.12

全局路由标头 (GRH)。

映射到本地路由器地址，目标路由器将通过将数据包的 DLID 更改为该子网中的最终目标端口来确保数据包被传递到目标节点。

11.5.4传输层传输层为上

层协议（ULP）（和应用程序）提供接口，以便使用 QP 进行发送和接收操作，通过网络层在子网内部和子网之间进行通信。它负责使用应用程序所需的传递特性（例如，可靠与不可靠以及连接与数据报）将数据有效负载从源端节点传递到目的地端节点。传输层根据传输头中的信息将数据包传送到正确的 QP。

传输层还负责向 ULP 提供分段和重新组装服务。它根据底层网络层支持的最大传输单元，将传输路径中的消费者数据分段为适当大小的有效负载。每个段在传输过程中都封装有报头和 CRC。接收后，QP 会重新组装内存中指定 ULP 缓冲区中的所有段。

数据的实际传输及其交付取决于给定 QP 配置的服务类型。有关这些机制的详细信息将在第 11.5.1 节中讨论。

11.6融合以太网上的 RDMA (RoCE)

11.6.1概述（DCB 和 RoCE）

人们越来越希望融合不同类型的结构、适配器，以降低总体 TCO（总拥有成本）。在单个物理基础设施上运行所有这些协议，而不是为 LAN、SAN（存储）和 IPC（低延迟）运行单独的网络，有很多好处。考虑到这一点，以太网标准得到了增强，以支持不同类型的网络（图 11.13）。

IEEE 定义了以太网的新增强功能，允许在 L2 网络中应用“无损”特性，这种增强型以太网称为 DCB（数据中心桥接）。

通过 DCB 网络，人们可以在以太网 L2 网络中获得类似于 IB 网络的“无损”特性。尽管这两种技术用于实现这种“无损”行为的机制有所不同，但出于所有实际目的，它们都实现了以无损方式通过链路传送数据包的类似结果（避免拥塞时丢失）。

由于 IB 协议设计为在无损第 2 层连接上运行，因此 DCB 提供了以太网中承载 IB 数据包所需的功能。

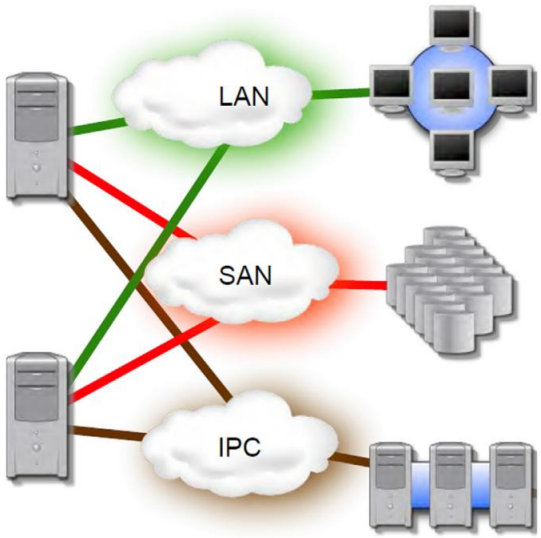


图11.13

I/O 融合。

RoCE 是一种通过以太网传输 IB 数据包的协议。它保持了大部分层的完整性，并使用以太网 L2 层作为物理层和链路层[9]
（图 11.14）。

11.6.2层架构从图11.14可以看出，
RoCE 维护了除链路层、MAC 层和物理层之外的所有层。

RoCE通过此次修改实现了以下目标：

1、采用DCB（无损以太网）作为二层网络，提供物理
连接性。
2. 使用 RDMA 作为 ULP 接口的应用程序没有变化
保持不变。
3. 维护现有的 IB 传输结构和服务（RC、UC、RD 等）
（图 11.15）。

11.6.3数据包格式

RoCE 将大部分 IB 数据包通过隧道传输到以太网数据包中[13]。以太网接头提供与 IB LRH 类似的功能。它允许以太网节点在给定子网中相互通信。因此，RoCE 数据包不包括隧道以太网数据包中的 LRH。 LRH 字段映射到等效的以太网标头字段。

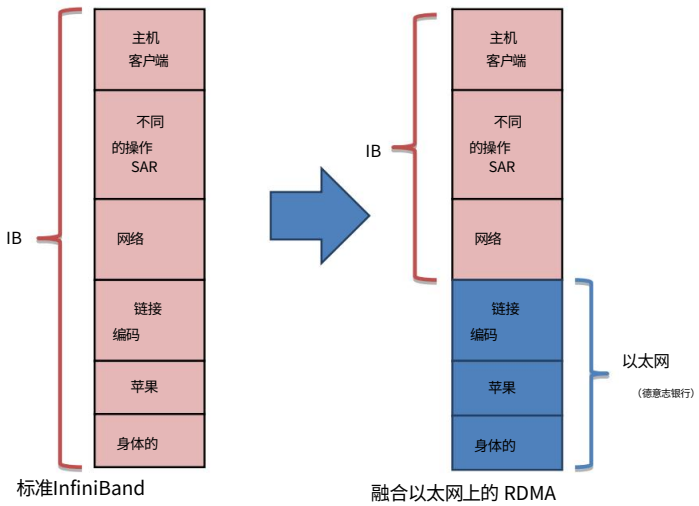


图11.14

IB和RoCE之间的层比较。

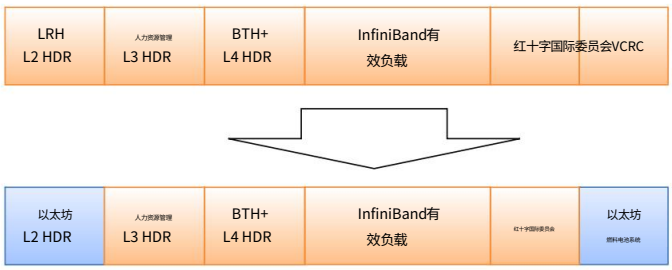


图11.15

IB和RoCE数据包格式比较。

由于以太网数据包覆盖有帧校验序列 (FCS),因此 RoCE 数据包中不需要来自 IB 数据包的 VCRC。 IB 数据包中的其余字段在 RoCE 数据包中完整携带。

11.6.4报头和地址映射以太网具有与 IB LRH 类似的报头结构（图 11.16）。

当 IB 数据包映射到以太网格式时,LRH 字段将替换为以太网 L2 标头。 DLID 和 SLID 被替换为 6 字节以太网 MAC 地址。 （IB 允许有关子网 LID 的信息

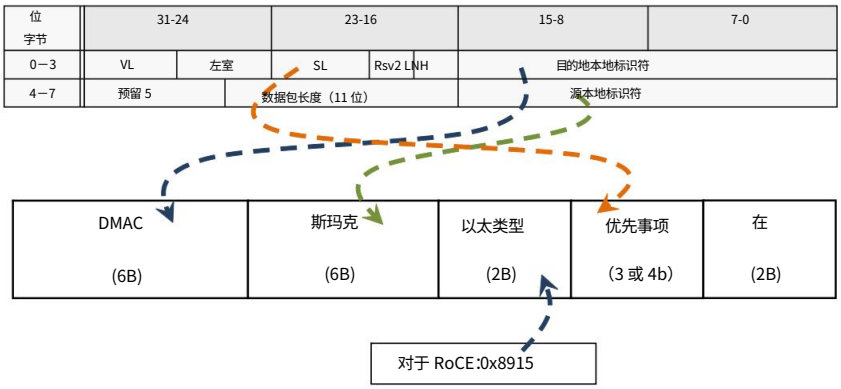


图11.16

RoCE 的 IB 标头映射。

ULP 通过动词接口访问。由于 RoCE 不携带 LRH,因此这些 LID 不会通过接口携带。）

MAC 地址通过普通以太网方法生成和使用（每个端点都有其分配的 MAC 地址）。 GID 与 MAC 地址的关联留给实现（通过类似于 ARP、邻居发现等众所周知的机制）

由于RoCE数据包中没有LRH,因此RoCE不支持原始服务。 （如第 11.5.3 节所示， Raw 服务不携带 GRH 和其他 IBA 标头,而是依赖于 LRH 标头;因此,RoCE 不支持它们）。

SL 在 VLAN 标头中的优先级/丢弃资格字段中表示。
由于 16 个 SL 有 8 个优先级值,因此 07 个 SL 值直接映射到 07 个优先级值。 SL 值 815 在 RoCE 中保留。

以太网标头没有像 VL 这样的字段来标识每个节点的本地资源（例如队列）。以太网标准允许通过编程接口将优先级值映射到本地队列（称为TC,流量类别）;但是,它没有一种机制可以在每个流/数据包上提供此类映射。因此,RoCE需要通过带外机制来实现这种映射。

RoCE 已分配 Ethertype (0 3 8915),用于识别以太网链路上的 RoCE 数据包。

11.6.5以太网结构要求RoCE 规范没有

明确要求 DCB 或“无损”以太网 然而,为了比较性能/功能,预计 RoCE 将仅与符合 DCB 的以太网交换机一起使用。

IEEE 802.1 于 2011 年定义了以下用于通过以太网提供融合流量的标准。

1. IEEE 802.1Qbb [14]:基于优先级的流量控制（PFC）
   A。 PFC 允许对以太网标头中用特定优先级位标识的流量进行选择性流量控制[7] b。提供以
   太网光纤通道 (FCoE) 和 RoCE 2 所需
   的无丢包行为。 IEEE 802.1Qaz:增强型传输选择 (ETS) ETS 为流量类别提供带宽
   分配;的替代品

   严格优先 B.
   DCBX 使用链路层发现协议 (LLDP) 协调跨链路的 DCB 功能配置

3. IEEE 802.1Qau:拥塞通知允许拥塞点通知流量源
   （反应点）
   拥塞

尽管大多数实现预计将转向这些标准,但市场上当前的实现遵循上述 #1 和 #2 的预标准多供应商协议规范[13]。

11.7 8 iWARP

11.7.1概述

互联网广域 RDMA 协议允许在以太网环境中通过 TCP/IP 使用 RDMA 协议。 iWARP 规范由 IETF（互联网工程任务组[1,46])标准化。 iWARP 为 ULP 提供了类似于 IB 和 RoCE 的动词接口。

11.7.2层架构IB 和 iWARP

由不同的标准机构定义,但它们都解决类似的网络需求,并为应用程序提供类似的动词接口。

图 11.17显示了 IB 和 iWARP 分层架构的大致比较。

图 11.18显示了 IETF 规范中定义的 iWARP 层。 iWARP 使用以太网在子网内进行本地路由,使用 IP 作为网络层来跨子网路由流量,使用 TCP 作为传输层来跨网络提供可靠且面向连接的数据包传输。 iWARP 还支持流控制传输协议 (SCTP) 作为另一种替代方案,作为远程直接内存访问协议 (RDMAP) 的传输层。 TCP 和 SCTP 之间的主要区别在于 TCP 是流协议（将消息转换为字节流）,而 SCTP 是面向消息的协议。两个都

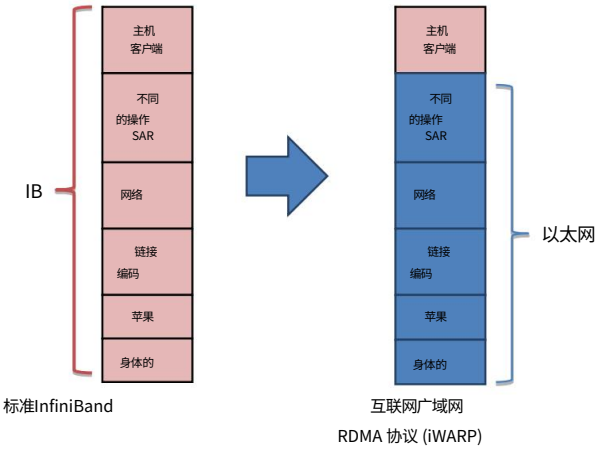


图11.17

IB 和 iWARP 之间的比较。

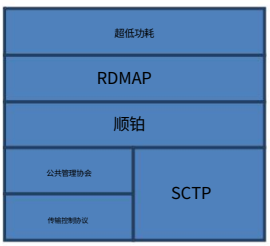


图11.18

iWARP 层架构。

协议在 IP 上运行,并且在拥塞管理机制方面彼此友好。 iWARP 为 TCP 和 SCTP 层上的 RDMA 提供 RC。目前,iWARP 在工业中最主要的用途是使用 TCP 作为传输层,因此本节的其余部分将重点介绍 TCP。

由于 iWARP 使用 TCP 作为传输层,因此它不依赖底层以太网结构来实现无损传输。由于 IP 可通过 Internet 进行路由,因此 iWARP 可在数据中心部署中进行路由（RoCE 尽管使用以太网结构,但不可在以太网数据中心中的 IP 子网之间进行路由）。

11.7.3数据包格式

11.7.3.1 RDMAP

RDMAP 是 RDMA 应用程序通过底层 iWARP 基础设施传输数据的接口。它使用底层直接数据替换（DDP)来

图11.19

iWARP 数据包标头。

使应用程序能够读取和写入远程节点的内存 (RDMA)
（图 11.19）。

11.7.3.2直接数据处理

DDP 可以将数据直接移动到目标节点（数据接收器)的内存中,而无需
要求网络接口将数据复制到中间缓冲区。这
层提供以下功能：

- 标记缓冲区模型:能够命名缓冲区并与其他人共享该信息
  对等点（这可以将数据直接放入目标节点的内存中）
- 无标记缓冲区模型:还允许将数据传输到匿名缓冲区
  数据接收器
- 可靠、按顺序交付
- ULP 消息的分段和重组;可以处理乱序
  不需要额外副本的片段

11.7.3.3 MPA（TCP 的标记 PDU 对齐成帧）

TCP 通过创建携带消息的段来跨网络传输消息
字节流。为了识别给定消息的边界
字节流,MPA已经定义。它放置边界标识符（标记）
在 TCP 流中允许接收者识别给定的边界
信息。

由于 SCTP 是面向消息的协议,因此在 SCTP 上运行 DDP 时不需要 MPA。 MPA 包括额外的 CRC 检查,以提高通过 TCP 运行时的数据完整性。

参考

[1] R. Recio、B. Metzler、P. Culley、J. Hilland、D. Garcia。远程直接存储器
访问协议规范[在线]。可从： http://tools.ietf.org/html/
RFC5040..

[2] CEE 规范 - DCBX,CEE 规范 - DCBX [在线]。可从：
,http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchange
discovery-protocol-1108-v1.01.pdf..

[3] CEE 规范 ETS [在线]。可从： http://www.ieee802.org/1/files/public/docs2008/az-wadekar-
dcbx-capability-exchange-discovery-protocol-1108-v1.01.pdf ..

[4] H. Shah、J. Pinkerton、R. Recio、P. Culley。通过可靠传输的直接数据放置[在线]。可从： http://tools.ietf.org/html/rfc5041..

[5] C. Bestler、R. Stewart,流控制传输协议 (SCTP) 直接数据放置 (DDP) 适配 [在线]。可从： http://tools.ietf.org/html/rfc5043 ..

[6] P. Culley、U. Elzur、R. Recio、S. Bailey、J. Carrier,TCP 规范的标记 PDU 对齐成帧 [在线]。可从： http://tools.ietf.org/html/rfc5044 ..

[7] CEE 规范 - PFC、CEE 规范 - PFC XE“PFC:基于优先级的流量控制”[在线]。可从： http://www.ieee802.org/1/files/public/docs2008/bb-pelissier-pfc-proposal-0508.pdf ..

[8] IBM,在 IBM Power Systems 上使用 InfiniBand 的 HPC 集群 [在线]。可用： http://www.redbooks.ibm.com/redbooks/pdfs/sg247767.pdf..

[9] 基于融合以太网的 RDMA:IB 架构规范第 1 卷版本 1.2.1 的补充 [在线]。可从： http://www.inifinibandta.org 获取。

[10] InfiniBand 架构规范第 1 卷和第 2 卷[在线]。可从： http://www.infinibandta.org 获取。

[11] GF Pfister [在线]。可从： http://gridbus.csse.unimelb.edu.au/Braj/超级存储/chap42.pdf..

[12] C. DeCusatis,光纤数据通信手册:光网络实用指南,第三版,学术出版社,2008 年。

[13] A. Ayoub,基于融合以太网的 RDMA (RoCE),2011 年 9 月 9 日 [在线]。可从以下网址获取： http://www.itc23.com/fileadmin/ITC23_files/slides/WDC_3_RoCE DC-CaVeS-9Sep2011-nb.pdf。

[14] IEEE 802.1Q，“数据中心桥接工作组”[在线]。可从： ,http://www.ieee802.org/1/pages/dcbridges.html..