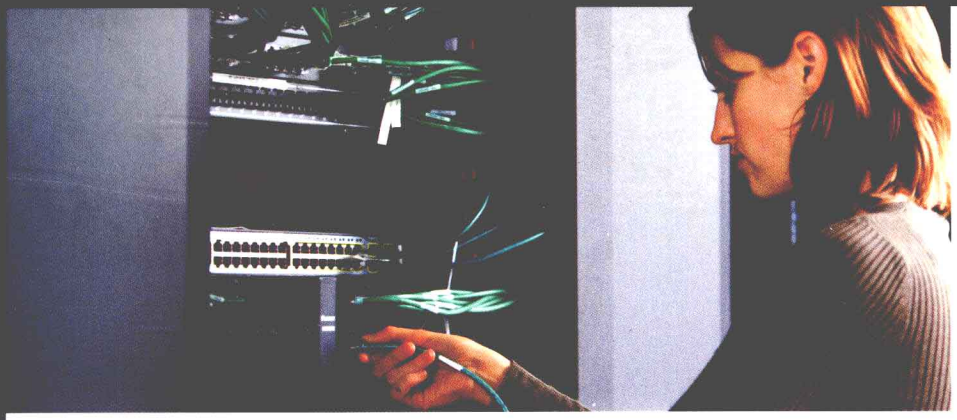
CISCO

ciscopress.com

思科数据中心系列



思科数据中心I/O整合

I/O Consolidation in the Data C[illegible]

A complete guide to D[illegible]
and Fibre Channel over Ethernet

〔美〕 Silvano Gai
Claudio DeSanti 著

陈柳 译

谭立勃 钱峻 曾维微 审校

人民邮电出版社
POSTS & TELECOM PRESS

ciscopress.com

思科数据中心系列

思科数据中心I/O整合

I/O Consolidation in the Data Center

〔美〕 Silvano Gai
Claudio DeSanti 著
陈柳 译
谭立勃 钱峻 曹维微 审校

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

思科数据中心I/O整合 / (美) 盖伊 (Gai, S.), (美) 德桑蒂 (DeSanti, C.) 著 ; 陈柳译. -- 北京 : 人民邮电出版社, 2013.1
ISBN 978-7-115-29210-0

Ⅰ. ①思… Ⅱ. ①盖… ②德… ③陈… Ⅲ. ①数据库系统一研究 Ⅳ. ①TP311.13

中国版本图书馆CIP数据核字(2012)第234831号

版权声明

思科数据中心 I/O 整合

◆ 著 [美] Silvano Gai Claudio DeSanti
译 陈 柳
审 校 谭立勃 钱 峻 曾维微
责任编辑 赵 轩

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街 14 号
邮编 100061 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn

北京艺辉印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：9.5
字数：175 千字 2013 年 1 月第 1 版
印数：1-3 000 册 2013 年 1 月北京第 1 次印刷
著作权合同登记号 图字：01-2012-5024 号

ISBN 978-7-115-29210-0

定价：49.00 元

读者服务热线：(010)67132692 印装质量热线：(010)67129223
反盗版热线：(010)67171154
广告经营许可证：京崇工商广字第 0021 号

目 录

第 1 章

I/O 整合

1.1 引言

在数据中心领域，目前以太网仍然是最主流的互连网络。最初的以太网只是作为一种共享媒体的技术，但是经过多年的发展，它已经成为了一种基于点对点、全双工链路的网络。现代数据中心主要部署百兆（100 Mbit/s）和千兆（1 Gbit/s）两种速率的以太网，适合当前基于 PCI（总线）I/O 性能的服务器。

存储流量明显是一个例外，因为它一般是通过基于光纤通道（FC）协议（群）而建立的一个独立网络进行传输的。大多数大型数据中心都会部署光纤通道，这些 FC 网络（也称为 Fabric）的规模一般并不大，不同的服务器组分别建有许多独立的 Fabric。大多数数据中心会出于高可用性的考虑而选择设置双重 FC Fabric。

高性能计算（HPC）领域和需要集群基础架构的应用程序，都会部署诸如 Myrinet 和 Quadrix 等专用和私有协议的网络。无限带宽技术（InfinityBand，简称 IB）已经进入了 HPC 领域和数据中心中特定的应用程序，且它能够很好地支持低延迟和高吞吐量（用户内存之间的访问）的集群。

图 1-1 所示为一个常见的数据中心配置，其包含一个以太网核心和为实现高可用性的两个独立 SAN Fabric（分别被标记为 SAN A 和 SAN B）。

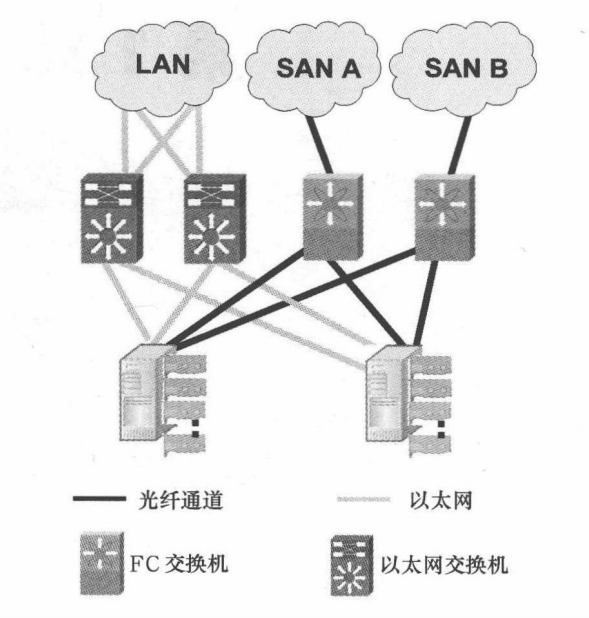


图 1–1：当前的数据中心架构

1.2　I/O 整合是什么

I/O 整合是指交换机或主机适配器能够使用同一个物理基础架构来传输多种类型的流量，而且每一种流量都具有独特的特性和特殊的处理需求。

从网络的方面考虑，这相当于只需要安装和运营一套网络，而非 3 套网络（参见图 1-2）。而在服务器和存储阵列方面，这相当于采用融合网络适配器（Converged Network Adapter，CNA）后，减少了以太网 NIC 卡、FC HBA 卡和 IB HCA 卡（的数量）。这样，服务器所需要的 PCI 插槽便减少了，对于刀片服务器而言，这是一个特别有利的情况。

这为客户带来的好处包括：

- 显著减少了布线，并简化和标准化了布线方式；
- 去除网关，它经常成为瓶颈并且会引起兼容性问题；
- 降低了对电源和制冷的需求；
- 降低了成本。

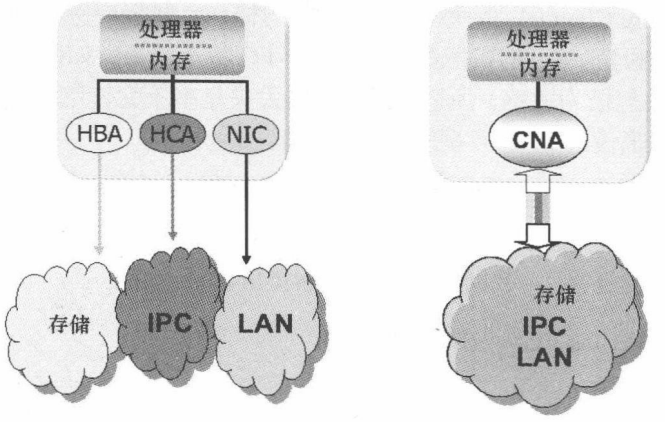


图 1–2：网络中的 I/O 整合

为了保证可行性，I/O 整合必须保持对当前各种类型的流量使用相同的管理理念（management paradigm）。

图 1-3 所示为用两个 CNA 替代两个 FC HBA 卡、两个以太网 NIC 卡和两个 IB HCA 卡的例子。

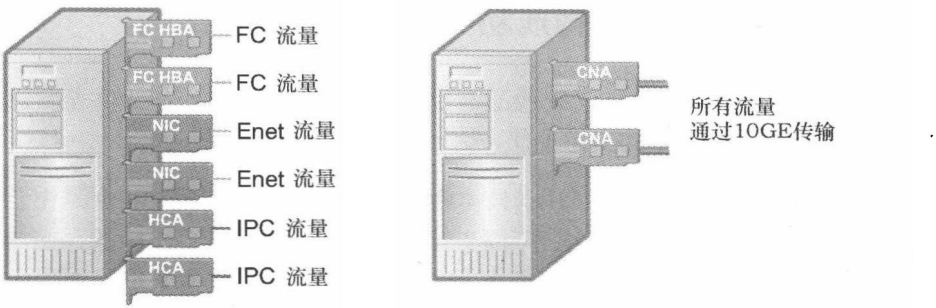


图 1–3：服务器中的 I/O 整合

1.3　整合的需求

I/O 整合的最大挑战在于如何在同一个网络中满足不同种类流量的需求。

目前，典型 LAN 流量主要由必须运行在原生以太网上的 IPv4 和 IPv6 通信协议组成。此点难以改变，因为这个领域里已经投入了巨大的投资，而且很多应用程序也将以太网假定为底层网络。这种通信的特点是包含大量的流。通常，这些流对延迟不敏感；但是此点也在快速演变，现在必须认真考虑延迟因素。流媒体流量对延迟抖动也非常敏感。

存储流量必须采用光纤通道（Fibre Channel，简称 FC）模型。同样，大型客户也在 FC 基础架构和管理上进行了大量的投资。存储开通通常会依赖于 FC 特定服务，如命名、分区等。因为 SCSI（协议）对丢包极为敏感，所以在 FC 中帧丢失是绝不允许的。FC 流量的特点是具有很多大包帧，来传输 2KB 典型的 SCSI 负载。

处理器间通信（Inter Processor Communication，简称 IPC）流量的特点是混合了大小不同的消息包/帧。它通常延迟敏感（尤其是短消息）。IPC 流量一般出现在集群中（例如，两台或两台以上计算机之间的连接）。数据中心的服务器集群例子包括：

- 可用性集群（例如 Symantec/Veritas VCS、微软 MSCS）；
- 群集文件系统；
- 群集数据库（例如，Oracle RAC）；
- VMware 虚拟基础架构服务（例如，VMware VMotion、VMware HA）。

如果成本较低，又拥有高带宽、低延迟，而且适配器还支持零拷贝（Zero-Copy）机制，那么集群就不太注重底层网络。

1.4 为什么 I/O 整合仍未取得成功

以前曾经出现过尝试实现 I/O 整合。光纤通道本身就曾经被提议作为一种 I/O 整合网络，但是由于它对多播/广播流量的支持不佳，因此一直未能得到认可。

在 HPC 领域的 I/O 整合中，无限带宽（IB）技术也曾经取得过一定的成功，但是由于它缺少对以太网（同样是没有很好的多播/广播支持）和 FC（使用与 FC 不同的存储协议）的兼容性，以及需要网关（可能引发瓶颈和兼容问题），所以它也未能获得更大的市场应用。

iSCSI 可能是 I/O 整合的一个最重要尝试。直到现在，它仍仅限于低端服务器，主要是因为以太网的最高速率为千兆（1 Gbit/s）。虽然万兆以太网（10 Gigabit Ethernet，10GE）突破了这个限制，但是万兆下 TCP 协议的负载（Overhead）是一个担心。真正的问题是，iSCSI 是“在 TCP 上实现的 SCSI”，而非“在 TCP 上实现的 FC”，因此它还无法保留 FC 的管理和部署模型。iSCSI 仍然需要网关，而且使用不同的命名模式（可能更好，但是不同）和分区方法（Zoning），等等。

1.5　基础技术

在实现 I/O 整合的过程中，有两种技术发挥着重要的作用，即 PCI-Express 和万兆以太网（10 Gigabit Ethernet，10GE）。

1.5.1　PCI–Express

PCI 是一种多年来广泛使用、支持计算机外围设备互连的旧标准。

PCI-Express（简称 PCI-E 或 PCIe）是一种专门用于替代 PCI、PCI-X 和 AGP 的计算机扩展卡接口格式。它突破了所有影响这些 I/O 整合方法的限制条件（例如，服务器总线端的 I/O 带宽不足），并且兼容当前的操作系统。

PCIe 使用称为线路（lane）的点对点全双工串行链路。每一条线路包含两对导线：一条用于发送；另一条用于接收。多条线路可以并行部署：1x 表示单条线路；4x 表示 4 条线路。

在 PCIe 1.1 中，线路速率为 2.5 Gbit/s（2 Gbit/s 的数据链路），支持 16 线路并行部署。支持的速度范围是 2 Gbit/s（1x）到 32 Gbit/s（16x）。由于协议的负载（overhead），支持 10 GE 网卡需要 8x。

PCIe 2.0（即 PCIe Gen 2）将每条线路的带宽翻倍，从 2 Gbit/s 变成 4 Gbit/s，最多线路数量可扩大到 32x，目前（基于它的产品）已经出货。

PCIe 3.0 计划将带宽增加近一倍："最终版 PCIe 3.0 标准包含了物理尺寸规格的更新，将会在 2009 年末发布，2010 年以后可能开始出现真正的产品。"

1.5.2　万兆以太网

从 2008 年开始，万兆以太网（简称 10GE）就已成为一种在实际中可行的互连技术，这个标准已经成熟，市场上也有廉价的布线解决方案。光纤继续用于远距离传输，而铜线则作为廉价线路部署在数据中心内部。

交换机和 CNA 是使用小型可插拔（Small Form-factor Pluggable，简称 SFP）收发器来实现接口的标准化。SFP 可以将网络设备的主板（例如，交换机、路由器或 CNA）连接到光纤或铜缆上。SFP 是一种受到多个零件供应商支持、流行的行业格式。它可以扩展成为 SFP+，最高支持 10 Gbit/s 速率的数据。SFP+的应用包括 8GFC 和 10GE。

SFP+的主要优点包括：

- 支持与 SFP 近似的面板密度；

- 模块功耗低于 XENPAK、X2 和 XFP；
- 标定 1W 功耗（可以选择 1.5W 高功耗模块）；
- 向下兼容 SFP 光学模块。

IEEE 双绞线电缆布线标准（10GBASE-T）目前还不是一种可行的互连技术，因为它需要大量的转发器（transistors），特别是在距离接近 100 米（328 英尺）时，这意味着电源功耗显著增加和额外的延迟时间（参见图 1-4）。设想一下，如何冷却一个交换机线卡，面板配备了 48 个 10GBASE-T 端口，每个端口功率为 4 瓦！

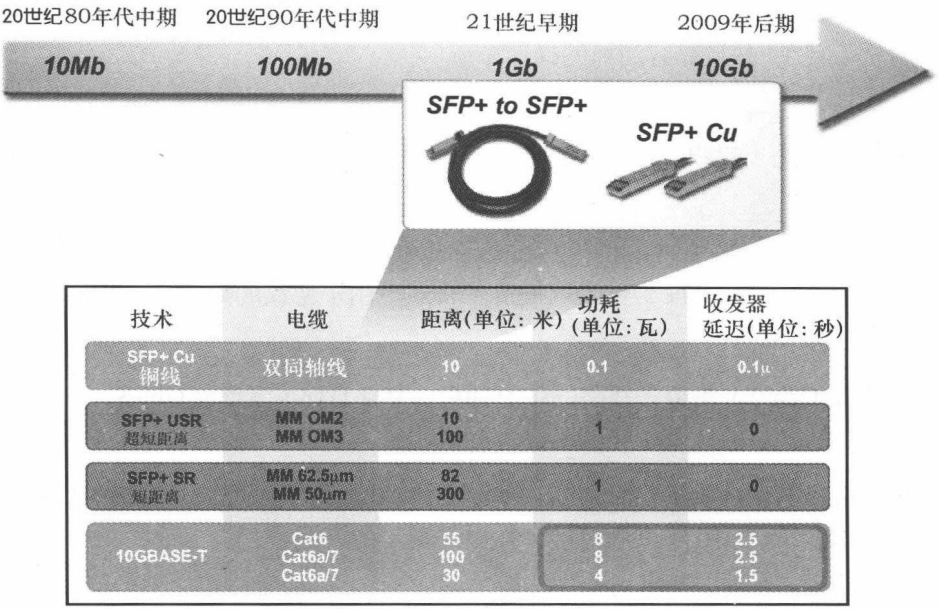


图 1-4：以太网物理介质演变

在数据中心机柜层面，更为实用的解决方案是使用 SFP+和双轴铜线（Twinax，由 SFF-8431 定义）。该导线很灵活，直径约 6 毫米（1/4 英寸），使用 SFP+作为接头。成本可以限制，功耗与延迟也可以忽略，并且它的长度被限制为 10 米（33 英尺），足以将少量机柜内的服务器与一个公用的架顶交换机进行连接。

思科、Amphenol、Molex、Panduit 等公司都提供此类线缆。

图 1-5 所示为在一个或多个机架中使用双轴线的优点。传输介质的成本，仅仅是在制造具有成本优势的 10GE 端口时需要考虑的因素之一，另外还需要考虑的因素有：交换缓存区（switch buffers）大小、2 层与 3/4 层协议上的功能差别等。

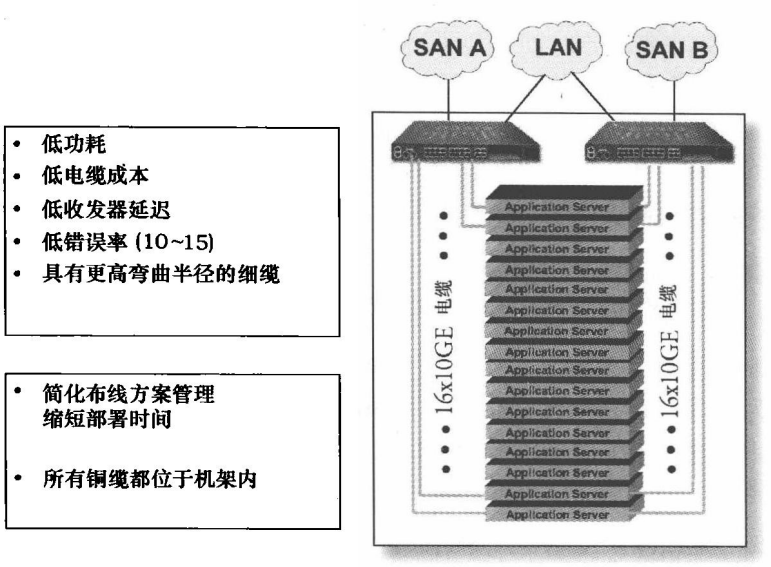


图 1–5：双轴铜线

1.6　其他需求

1.6.1　缓存需求

缓存是一个复杂的问题，涉及传播延迟、更高层协议、拥塞控制模式等。按照目前所讨论的内容，可以将网络划分为两类：无损耗网络和有损耗网络。

在可控和短距离传输环境中（如数据中心），这种分类方式并不考虑传输错误导致的损耗。因为相对于拥塞造成的损耗，传输错误导致的损耗是微不足道的。

光纤通道（FC）和无限宽带（IB）技术就是无损耗网络（例如，它们使用链路层信令机制跟踪链路另一端的缓存区可用性）。这种机制允许发送者在接收端缓存区可用时才发送数据帧，因此接收端不会出现丢帧情况。虽然这种方法初看起来很吸引人，但是一定要注意：无损耗网络需要针对简单和有限的拓扑结构进行重设计。事实上，交换机拥塞可能会从上游网络漫延到整个网络，最终会影响到与拥塞无关的网络流。如果出现循环依赖，网络可能会出现严重的死锁和/或活锁状况，进而显著影响网络性能或者可能破坏网络功能。上述这两种现象都在文献中广为人知，也很容易复现；但这应该不会影响潜在的用户，因为数据中心网络拥有简单且设计完备的网络拓扑结构。

历史上，以太网一直是一种有损耗网络，因为以太网交换机没有使用任何机制提醒发送端是否出现缓存区耗尽的问题。几年前，IEEE 802.3 标准给以太网增加了 PAUSE

机制。这种机制可以将发送端中断一定时间，但是，实际上这个特性并未能成功部署。现在，当以太网交换机出现拥塞状况，经常会出现丢帧的现象。主动队列管理（Active Queue Management，简称 AQM）已经提出了几种处理丢帧和管理队列的方法，但是它们仍然未能解决丢帧的问题，并且需要使用较大的缓存区。最常用的 AQM 模式可能就是随机早期检测（Random Early Detection，RED）。

通过以太网传输原生存储流量必须解决丢帧问题，因为存储流量不允许丢帧。SCSI 在设计上假设通过可靠介质进行传输，由于其出错率极低，因此可以接受缓慢的恢复。

光纤通道是用于传输存储流量的主要协议，它通过一种链路流控制机制解决丢帧问题，该机制基于一种名为缓存区间流控制（Buffer-to-buffer control）（也称为缓存区间信用或 B2B 信用）的信用机制而得以实现。iSCSI 是光纤通道的一种替代方法，它通过要求 TCP 恢复丢失的帧来解决此问题；然而，iSCSI 还未能在数据中心内广泛部署。

一般而言，无损耗网络的交换机缓存区需求小于有损耗网络，这些缓存区既可以在芯片上实现（更低廉和更快速），也可以通过芯片外的内存（昂贵且速度慢）实现（通常用于满足大缓存的要求）。

这两种行为各具优点和缺点。以太网需要扩展功能，将物理链路划分为多个本地链路（通过扩展 IEEE 802.1Q 优先级概念），允许按照划分的优化级规定无损耗/有损耗行为。

最后，值得注意的是：使用缓存区会增加延迟时间。

1.6.2 只支持 2 层协议

在 10GE 交换机内的端口（inter-switch port）生产成本中，主要和 2 层以上的协议功能密切相关，例如 IPv4/IPv6 路由选择、多播转发、各种通道协议、多协议标记交换（MPLS）、访问控制列表（ACL）和深度包检测（4 层及以上协议）。这些特性需要外部组件提供的支持，如 RAM、CAM 或 TCAM，这些都会显著增加端口成本。

虚拟化、集群和 HPC 通常对 2 层协议连接有极高的要求。虚拟机通常在同一个 IP 子网中进行移动（2 层协议域），通常会使用无偿 ARP（gratuitous ARP）等 2 层协议。集群成员之间会交换大量的数据，通常还使用非 IP 协议实现成员关系管理、PING 和保持连接等功能。

如果某个 10GE 解决方案具有线速、低延迟和完全兼容以太网协议等特点，即使无法将它扩展到数据中心外部，在数据中心内部署也很适合。规模在 64000 到 256000 的二层协议域能够满足数据中心未来几年的需求。

为了在同一个网络中支持多个独立的流量类型，一定要保持 VLAN 概念，同时需要扩

充优先级概念。

1.6.3 交换架构

这一节将介绍存储转发与直通式交换两者之间一直存在的争论。许多读者可能会很反感，因为他们已经听到太多诸如此类的争论了，而且某些厂商也在这两个方向上摇摆不定，并且读者的反应是对的！

在以太网速率较慢时（例如，10Mbit/s 或 100Mbit/s），存储转发阵营理所当然占据上风，因为序列化延迟是主导原因。现在，已经有了 10GE，并且 40GE 和 100GE 也将随之相继出现，序列化延迟已经低到值得对这个问题进行重新考虑。例如，在 10Gbit/s 的速率下，1KB 数据帧只需要 1 毫秒左右时间就能够完成序列化。

现在，许多以太网交换机在设计上都采用存储转发架构，因为这种设计更为简单。存储转发会在交换机内部增加一定的延迟，从而会对总体延迟造成负面影响。

直通式交换机通过复杂的设计，带来了延迟时间的降低，该设计减少了中间的存储转发时间，这可以在 Nexus 5000 等固定（端口）配置交换机上实现，但是在 Nexus 7000 等具有大量端口的模块化交换机上，则会存在较多的问题。

在固定（端口）配置的交换机上，所有组件设计都使用相同的速率（例如，10Gbit/s），选择的端口数量也有限制（一般小于 128 个），这些简化后的前提条件都增加了实现直通式交换的可能性。

在模块化交换机中，存在多个背板交换 Fabric（目的在于改进高可用性、模块化和可服务性），并且采用尽可能高速的链路与线卡连接。模块化交换机可能拥有成千上万个端口，因为它们拥有大数量的线卡，而每一条线卡又拥有大量的端口。线卡种类各不相同（1GE、10GE、40GE 等），而前面板端口的速度又低于背板（Fabric）端口的速度。因此，在入口线卡和 Fabric 之间，以及在 Fabric 与出口线卡之间存在两次存储转发是不可避免的。

如果特定数据帧已经进入排队帧，而且出口链路速度高于入口链路（Data Underrun，则直通式交换是不可能实现的。针对多播/广播帧一般也不能进行直通式交换。

最后，直通式交换机无法丢弃损坏的帧，因为当通过检查帧控制序列（Frame Control Sequence，简称 FCS）发现错误帧时，这些帧已经开始传输了。

1.6.4 低延迟

集群用户关心的延迟参数是，将一台计算机用户内存空间中的缓存数据传输到另一台

计算机用户内存时发生的延迟。导致延迟的主要因素有以下几个方面。

（1）应用程序提交数据时间到与第一个字节开始在线路传输的时间差：即使这些内容是在分散的物理内存中，这个时间差也仍然取决于零拷贝机制（Zero-copy mechanism）和网卡直接访问服务器物理内存的能力。为了将这个时间保持在限定范围之内，现在大多数 NIC 都会使用 DMA 分散/收集操作指令，在内存与 NIC 之间传输帧。反过来，这也受到所使用的协议卸载类型的影响（例如，无状态还是 TOE［TCP 卸载引擎］）。

（2）序列化延迟：只取决于链路速度。例如，在 10Gbit/s 速率下，1KB 数据的序列化需要 0.8 毫秒。

（3）传播延迟：铜线和光纤的速度类似，一般能达到光速的 2/3，单程速度为 200 米/毫秒或者双程 100 米/毫秒；有人也会将其描述为 5 纳秒/米，这种说法也是正确的。在公布的延迟数据中，传输延迟往往被假定为 0。数据中心网络的规模必须被限定为几百米，这样才能将延迟时间保持在较低水平，否则会使延迟大幅提高，无法达到低延迟的要求。

（4）拥塞或无拥塞状况下的交换机延迟：在拥塞情况下，交换机延迟时间主要受到交换机内部的缓存影响，因此无法实现低延迟要求；在无拥塞状况下，延迟时间主要取决于交换机架构。

（5）与第 1 点相同，但是特指接收端。

1.6.5 存储流量的原生支持

所谓存储流量的原生支持，是指网络提供 SCSI 协议传输的能力。图 1-6 说明了 SCSI 传输的各种替代方法。

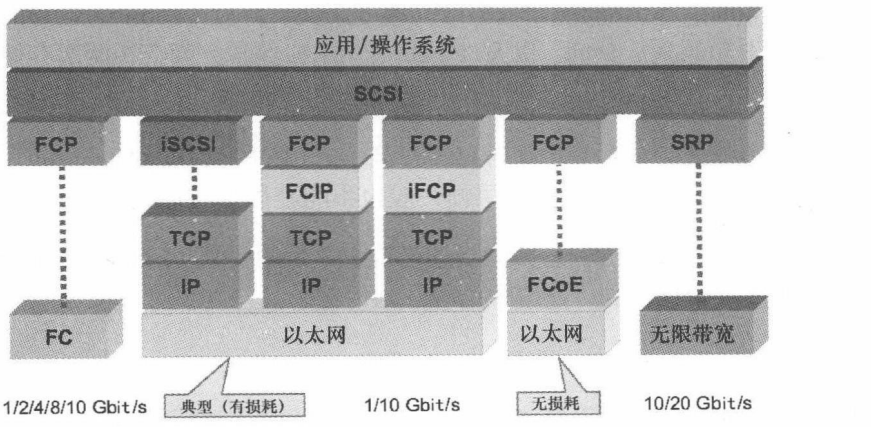


图 1-6：SCSI 传输