机的任务是接收入链路层帧并将它们转发到出链路；我们将在这一节中详细学习这种转发功能。我们将看到交换机自身对子网中的主机和路由器是**透明的**（transparent）；这就是说，某主机/路由器向另一个主机/路由器寻址一个帧（而不是向交换机寻址该帧），顺利地将该帧发送进局域网，并不知道某交换机将会接收该帧并将它转发到另一个节点。这些帧到达该交换机的任何输出接口之一的速率可能暂时会超过该接口的链路容量。为了解决这个问题，交换机输出接口设有缓存，这非常类似于路由器接口为数据报设有缓存。现在我们来仔细考察交换机运行的原理。

1. 交换机转发和过滤

过滤（filtering）是决定一个帧应该转发到某个接口还是应当将其丢弃的交换机功能。**转发**（forwarding）是决定一个帧应该被导向哪个接口，并把该帧移动到那些接口的交换机功能。交换机的过滤和转发借助于**交换机表**（switch table）完成。该交换机表包含某局域网上某些主机和路由器的但不必是全部的表项。交换机表中的一个表项包含：①一个MAC 地址；②通向该 MAC 地址的交换机接口；③表项放置在表中的时间。图 6-22 中显示了图 6-15 中最上方交换机的一个交换机表的例子。尽管帧转发的描述听起来类似于第 4 章讨论的数据转发，但我们将很快看到它们之间有重要的差异。确实在 4.4 节中一般化转发的讨论中我们学习过，许多现代分组交换机能够被配置，以基于第二层目的 MAC 地址（即起着第二层交换机的功能）或者第三层 IP 目的地址（即起着第三层交换机的功能）进行转发。无论如何，我们将对交换机基于 MAC 地址而不是基于 IP 地址转发分组进行明确区分。我们也将看到传统的（即处于非 SDN 环境）交换机表的构造方式与路由器转发表的构造方式有很大不同。

地址	接口	时间
62-FE-F7-11-89-A3	1	9:32
7C-BA-B2-B4-91-10	3	9:36
…	…	…

图 6-22　图 6-15 中最上面交换机的交换机表的一部分

为了理解交换机过滤和转发的工作过程，假定目的地址为 DD-DD-DD-DD-DD-DD 的帧从交换机接口 x 到达。交换机用 MAC 地址 DD-DD-DD-DD-DD-DD 索引它的表。有 3 种可能的情况：

- 表中没有对于 DD-DD-DD-DD-DD-DD 的表项。在这种情况下，交换机向除接口 x 外的所有接口前面的输出缓存转发该帧的副本。换言之，如果没有对于目的地址的表项，交换机广播该帧。
- 表中有一个表项将 DD-DD-DD-DD-DD-DD 与接口 x 联系起来。在这种情况下，该帧从包括适配器 DD-DD-DD-DD-DD-DD 的局域网网段到来。无须将该帧转发到任何其他接口，交换机通过丢弃该帧执行过滤功能即可。
- 表中有一个表项将 DD-DD-DD-DD-DD-DD 与接口 $y \neq x$ 联系起来。在这种情况下，该帧需要被转发到与接口 y 相连的局域网网段。交换机通过将该帧放到接口 y 前面的输出缓存完成转发功能。

我们大致地看一下用于图 6-15 中最上面交换机的这些规则以及图 6-22 中所示的它的交换机表。假设目的地址为 62-FE-F7-11-89-A3 的一个帧从接口 1 到达该交换机。交换机检查它的表并且发现其目的地是在与接口 1 相连的局域网网段上（即电气工程系的局域

网）。这意味着该帧已经在包含目的地的局域网网段广播过了。因此该交换机过滤（即丢弃）了该帧。现在假设有同样目的地址的帧从接口2到达。交换机再次检查它的表并且发现其目的地址在接口1的方向上；因此它向接口1前面的输出缓存转发该帧。这个例子清楚地表明，只要交换机的表是完整和准确的，该交换机无须任何广播就向着目的地转发帧。

在这种意义上，交换机比集线器更为“聪明”。但是一开始这个交换机表是如何配置起来的呢？链路层有与网络层路由选择协议等价的协议吗？或者必须要一名超负荷工作的管理员人工地配置交换机表吗？

2. 自学习

交换机具有令人惊奇的特性（特别是对于早已超负荷工作的网络管理员），那就是它的表是自动、动态和自治地建立的，即没有来自网络管理员或来自配置协议的任何干预。换句话说，交换机是**自学习**（self-learning）的。这种能力是以如下方式实现的：

1）交换机表初始为空。

2）对于在每个接口接收到的每个入帧，该交换机在其表中存储：①在该帧源地址字段中的MAC地址；②该帧到达的接口；③当前时间。交换机以这种方式在它的表中记录了发送节点所在的局域网网段。如果在局域网上的每个主机最终都发送了一个帧，则每个主机最终将在这张表中留有记录。

3）如果在一段时间（称为*老化期*（aging time））后，交换机没有接收到以该地址作为源地址的帧，就在表中删除这个地址。以这种方式，如果一台PC被另一台PC（具有不同的适配器）代替，原来PC的MAC地址将最终从该交换机表中被清除掉。

我们粗略地看一下用于图6-15中最上面交换机的自学习性质以及在图6-22中它对应的交换机表。假设在时刻9:39，源地址为01-12-23-34-45-56的一个帧从接口2到达。假设这个地址不在交换机表中。于是交换机在其表中增加一个新的表项，如图6-23中所示。

地址	接口	时间
01-12-23-34-45-56	2	9:39
62-FE-F7-11-89-A3	1	9:32
7C-BA-B2-B4-91-10	3	9:36
…	…	…

图6-23　交换机学习到地址为01-12-23-34-45-56的适配器所在的位置

继续这个例子，假设该交换机的老化期是60min，在9:32～10:32期间源地址是62-FE-F7-11-89-A3的帧没有到达该交换机。那么在时刻10:32，这台交换机将从它的表中删除该地址。

交换机是**即插即用设备**（plug-and-play device），因为它们不需要网络管理员或用户的干预。要安装交换机的网络管理员除了将局域网网段与交换机的接口相连外，不需要做其他任何事。管理员在安装交换机或者当某主机从局域网网段之一被去除时，他没有必要配置交换机表。交换机也是双工的，这意味着任何交换机接口能够同时发送和接收。

3. 链路层交换机的性质

在描述了链路层交换机的基本操作之后，我们现在来考虑交换机的特色和性质。我们能够指出使用交换机的几个优点，它们不同于如总线或基于集线器的星形拓扑那样的广播链路：

- 消除碰撞。在使用交换机（不使用集线器）构建的局域网中，没有因碰撞而浪费

的带宽！交换机缓存帧并且决不会在网段上同时传输多于一个帧。就像使用路由器一样，交换机的最大聚合带宽是该交换机所有接口速率之和。因此，交换机提供了比使用广播链路的局域网高得多的性能改善。

- 异质的链路。交换机将链路彼此隔离，因此局域网中的不同链路能够以不同的速率运行并且能够在不同的媒体上运行。例如，图 6-22 中最上面的交换机有 3 条 1Gbps 1000BASE-T 铜缆链路、2 条 100Mbps 10BASE-FX 光缆链路和 1 条 100BASE-T 铜缆链路。因此，对于原有的设备与新设备混用，交换机是理想的。
- 管理。除了提供强化的安全性（参见插入材料“关注安全性”），交换机也易于进行网络管理。例如，如果一个适配器工作异常并持续发送以太网帧（称为快而含糊的（jabbering）适配器），交换机能够检测到该问题，并在内部断开异常适配器。有了这种特色，网络管理员不用起床并开车到工作场所去解决这个问题。类似地，一条割断的缆线仅使得使用该条缆线连接到交换机的主机断开连接。在使用同轴电缆的时代，许多网络管理员花费几个小时“沿线巡检”（或者更准确地说“在天花板上爬行”），以找到使整个网络瘫痪的电缆断开之处。交换机也收集带宽使用的统计数据、碰撞率和流量类型，并使这些信息为网络管理者使用。这些信息能够用于调试和解决问题，并规划该局域网在未来应当演化的方式。研究人员还在原型系统部署中探讨在以太局域网中增加更多的管理功能［Casado 2007；Koponen 2011］。

关注安全性

嗅探交换局域网：交换机毒化

当一台主机与某交换机相连时，它通常仅接收到明确发送给它的帧。例如，考虑在图 6-17 中的一个交换局域网。当主机 A 向主机 B 发送帧时，在交换机表中有用于主机 B 的表项，则该交换机将仅向主机 B 转发该帧。如果主机 C 恰好在运行嗅探器，主机 C 将不能够嗅探到 A 到 B 的帧。因此，在交换局域网的环境中（与如 802.11 局域网或基于集线器的以太局域网的广播链路环境形成对比），攻击者嗅探帧更为困难。然而，因为交换机广播那些目的地址不在交换机表中的帧，位于 C 上的嗅探器仍然能嗅探某些不是明确寻址到 C 的帧。此外，嗅探器将能够嗅探到具有广播地址 FF-FF-FF-FF-FF-FF 的广播帧。一个众所周知的对抗交换机的攻击称为**交换机毒化**（switch poisoning），它向交换机发送大量的具有不同伪造源 MAC 地址的分组，因而用伪造表项填满了交换机表，没有为合法主机留下空间。这使该交换机广播大多数帧，这些帧则能够由嗅探器俘获到［Skoudis 2006］。由于这种攻击只有技艺高超的攻击者才能做到，因此交换机比起集线器和无线局域网来更难受到嗅探。

4. 交换机和路由器比较

如我们在第 4 章学习的那样，路由器是使用网络层地址转发分组的存储转发分组交换机。尽管交换机也是一个存储转发分组交换机，但它和路由器是根本不同的，因为它用 MAC 地址转发分组。交换机是第二层的分组交换机，而路由器是第三层的分组交换机。然而，回顾我们在 4.4 节中所学习的内容，使用“匹配加动作”的现代交换机能够转发基

于帧的目的 MAC 地址的第二层帧，也能转发使用数据报目的 IP 地址的第三层数据报。我们的确看到了使用 OpenFlow 标准的交换机能够基于 11 个不同的帧、数据报和运输层首部字段，执行通用的分组转发。

即使交换机和路由器从根本上是不同的，网络管理员在安装互联设备时也经常必须在它们之间进行选择。例如，对于图 6-15 中的网络，网络管理员本来可以很容易地使用路由器而不是交换机来互联各个系的局域网、服务器和互联网网关路由器。路由器的确使得各系之间通信而不产生碰撞。既然交换机和路由器都是候选的互联设备，那么这两种方式的优点和缺点各是什么呢?

首先考虑交换机的优点和缺点。如上面提到的那样，交换机是即插即用的，这是世界上所有超负荷工作的网络管理员都喜爱的特性。交换机还能够具有相对高的分组过滤和转发速率，就像图 6-24 中所示的那样，交换机必须处理高至第二层的帧，而路由器必须处理高至第三层的数据报。在另一方面，为了防止广播帧的循环，交换网络的活跃拓扑限制为一棵生成树。另外，一个大型交换网络将要求在主机和路由器中有大的 ARP 表，这将生成可观的 ARP 流量和处理量。而且，交换机对于广播风暴并不提供任何保护措施，即如果某主机出了故障并传输出没完没了的以太网广播帧流，该交换机将转发所有这些帧，使得整个以太网的崩溃。

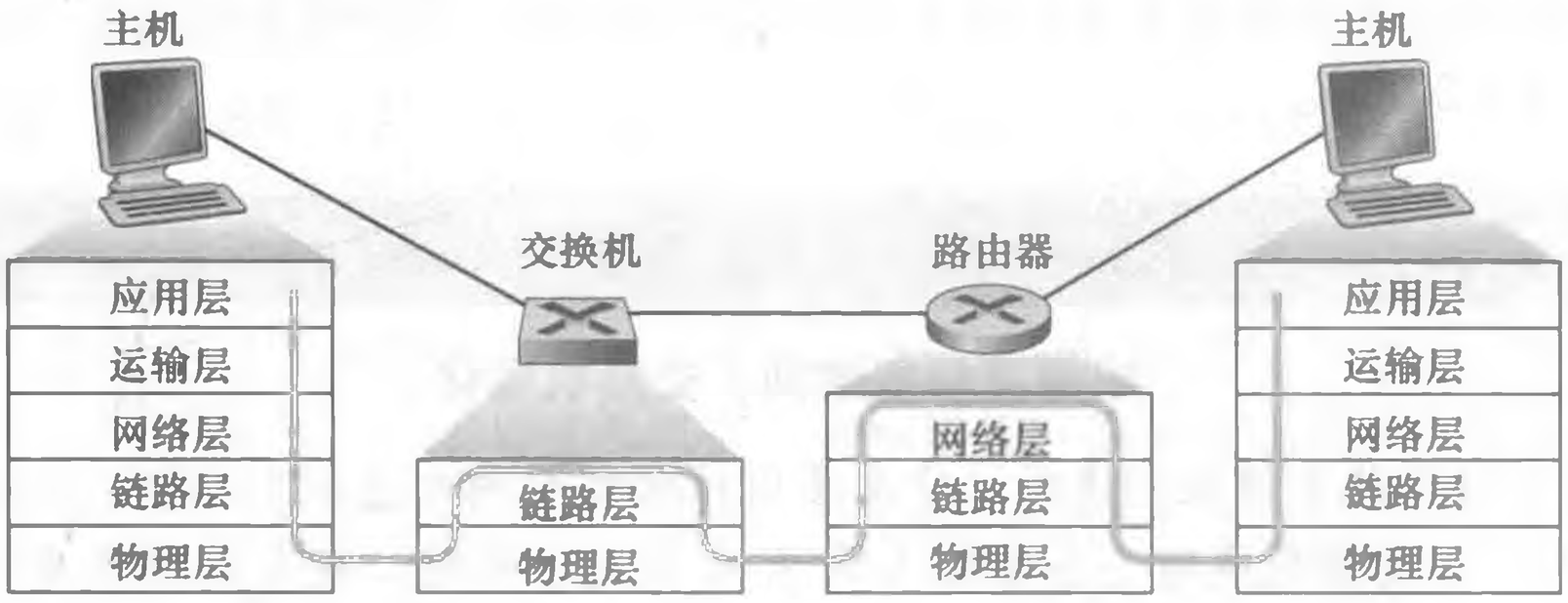


图 6-24　在交换机、路由器和主机中分组的处理

现在考虑路由器的优点和缺点。因为网络寻址通常是分层次的（不像 MAC 寻址那样是扁平的），即使当网络中存在冗余路径时，分组通常也不会通过路由器循环。（然而，当路由器表被误配置时，分组可能循环；但是如我们在第 4 章所知，IP 用一个特殊的报文首部字段来限制循环。）所以，分组就不会被限制到一棵生成树上，并可以使用源和目的地之间的最佳路径。因为路由器没有生成树限制，所以它们允许以丰富的拓扑结构构建因特网，例如包括欧洲和北美之间的多条活跃链路。路由器的另一个特色是它们对第二层的广播风暴提供了防火墙保护。尽管也许路由器最重要的缺点就是它们不是即插即用的，即路由器和连接到它们的主机都需要人为地配置 IP 地址。而且路由器对每个分组的处理时间通常比交换机更长，因为它们必须处理高达第三层的字段。最后，路由器一词有两种不同的发音方法，或者发音为“rootor”或发音为“rowter”，人们浪费了许多时间争论正确的发音［Perlman 1999］。

给出了交换机和路由器各自具有的优点和缺点后（总结在表 6-1 中），一个机构的网络（例如，大学校园网或者公司园区网）什么时候应该使用交换机，什么时候应该使

表 6-1　流行的互联设备的典型特色的比较

	集线器	路由器	交换机
流量隔离	无	有	有
即插即用	有	无	有
优化路由	无	有	无

用路由器呢？通常，由几百台主机组成的小网络通常有几个局域网网段。对于这些小网络，交换机就足够了，因为它们不要求 IP 地址的任何配置就能使流量局部化并增加总计吞吐量。但是在由几千台主机组成的更大网络中，通常在网络中（除了交换机之外）还包括路由器。路由器提供了更健壮的流量隔离方式和对广播风暴的控制，并在网络的主机之间使用更“智能的”路由。

对于交换网络和路由网络的优缺点的进一步讨论，以及如何能够将交换局域网技术扩展为比今天的以太网容纳多两个数量级以上的主机，参见［Meyers 2004；Kim 2008］。

6.4.4　虚拟局域网

在前面图 6-15 的讨论中，我们注意到现代机构的局域网常常是配置为等级结构的，每个工作组（部门）有自己的交换局域网，经过一个交换机等级结构与其他工作组的交换局域网互联。虽然这样的配置在理想世界中能够很好地工作，但在现实世界常常不尽如人意。在图 6-15 中的配置中，能够发现 3 个缺点：

- 缺乏流量隔离。尽管该等级结构把组流量局域化到一个单一交换机中，但广播流量（例如携带 ARP 和 DHCP 报文或那些目的地还没有被自学习交换机学习到的帧）仍然必须跨越整个机构网络。限制这些广播流量的范围将改善局域网的性能。也许更为重要的是，为了安全/隐私的目的也可能希望限制局域网广播流量。例如，如果一个组包括公司的行政管理团队，另一个组包括运行着 Wireshark 分组嗅探器的心怀不满的雇员，网络管理员也许非常希望行政流量无法到达该雇员的主机。通过用路由器代替图 6-15 中的中心交换机，能够提供这种类型的隔离。我们很快看到这种隔离也能够经过一种交换（第二层）解决方案来取得。
- 交换机的无效使用。如果该机构不止有 3 个组，而是有 10 个组，则将要求有 10 个第一级交换机。如果每个组都较小，比如说少于 10 个人，则单台 96 端口的交换机将足以容纳每个人，但这台单一的交换机将不能提供流量隔离。
- 管理用户。如果一个雇员在不同组间移动，必须改变物理布线，以将该雇员连接到图 6-15 中的不同的交换机上。属于两个组的雇员将使问题更为困难。

幸运的是，这些难题中的每个都能够通过支持**虚拟局域网**（Virtula Local Network，VLAN）的交换机来处理。顾名思义，支持 VLAN 的交换机允许经一个单一的物理局域网基础设施定义多个虚拟局域网。在一个 VLAN 内的主机彼此通信，仿佛它们（并且没有其他主机）与交换机连接。在一个基于端口的 VLAN 中，交换机的端口（接口）由网络管理员划分为组。每个组构成一个 VLAN，在每个 VLAN 中的端口形成一个广播域（即来自一个端口的广播流量仅能到达该组中的其他端口）。图 6-25 显示了具有 16 个端口的单一交换机。端口 2～8 属于电气工程系（EE）VLAN，而端口 9～15 属于计算机科学系（CS）VLAN（端口 1 和 16 未分配）。这个 VLAN 解决了上面提到的所有困难，即 EE VLAN 帧和 CS VLAN 帧彼此隔离，图 6-15 中的两台交换机已由一台交换机替代，并且在交换机端口 8 的用户加入计算机科学系时，网络操作员只需重新配置 VLAN 软件，

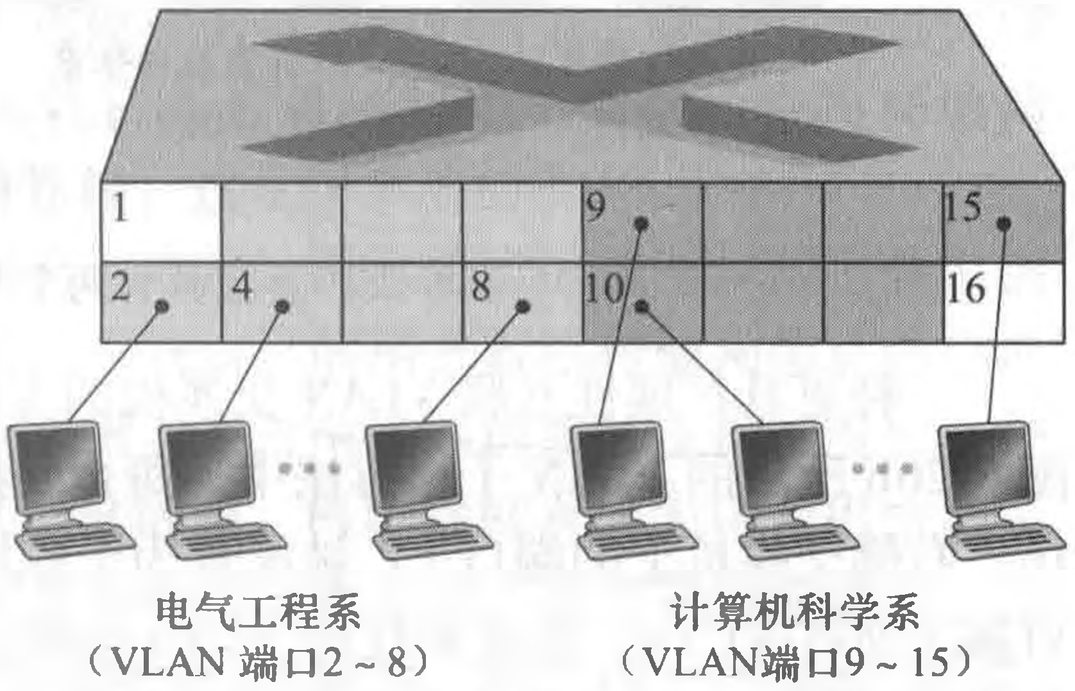
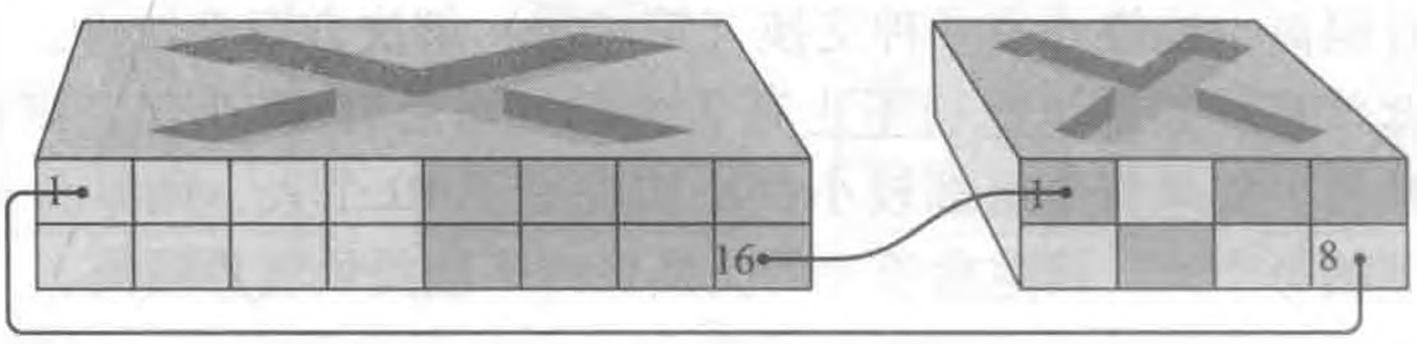


图 6-25　配置了两个 VLAN 的单台交换机

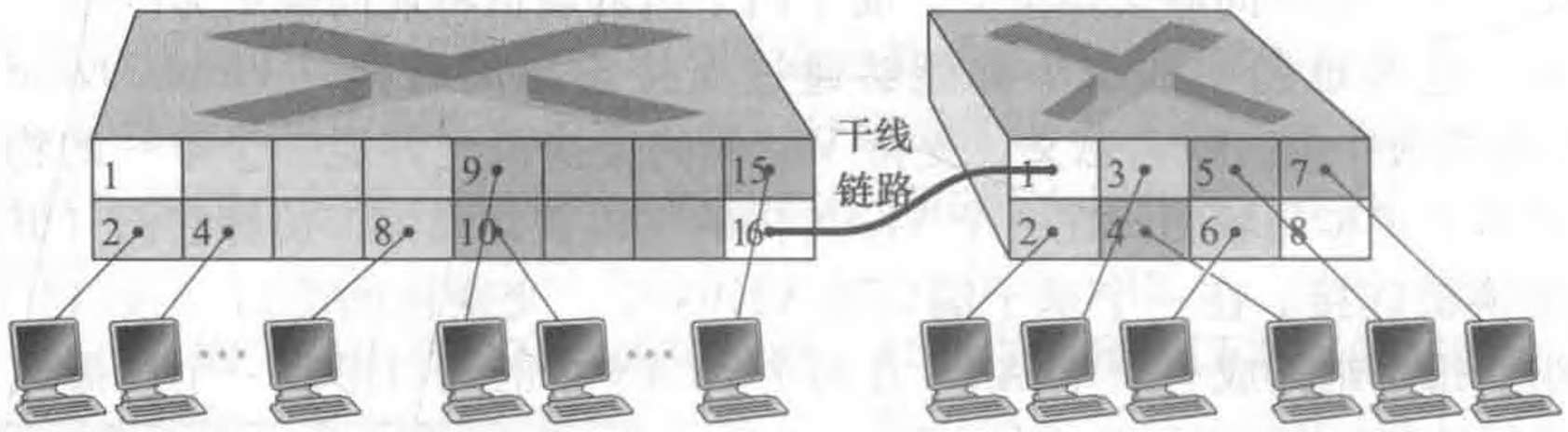
使得端口8与CS VLAN相关联即可。人们容易想象到VLAN交换机配置和操作的方法，即网络管理员使用交换机管理软件声明一个端口属于某个给定的VLAN（其中未声明的端口属于一个默认的VLAN），在交换机中维护一张端口到VLAN的映射表；交换机软件仅在属于相同VLAN的端口之间交付帧。

但完全隔离两个VLAN带来了新的困难！来自电子工程系的流量怎样才能发送到计算机科学系呢？解决这个问题的一种方式是将VLAN交换机的一个端口（例如在图6-25中的端口1）与一台外部的路由器相连，并且将该端口配置为属于EE VLAN和CS VLAN。在此情况下，即使电子工程系和计算机科学系共享相同的物理交换机，其逻辑配置看起来也仿佛是电子工程系和计算机科学系具有分离的经路由器连接的交换机。从电子工程系发往计算机科学系的数据报将首先跨越EE VLAN到达路由器，然后由该路由器转发跨越CS VLAN到达CS主机。幸运的是交换机厂商使这种配置变得容易，网络管理员通过构建包含一台VLAN交换机和一台路由器的单一设备，这样就不再需要分离的外部路由器了。本章后面的课后习题中更为详细地探讨了这种情况。

再次返回到图6-15，我们现在假设计算机工程系没有分离开来，某些电子工程和计算机科学教职员工位于一座建筑物中，他们当然需要网络接入，并且他们希望成为他们系VLAN的一部分。图6-26显示了第二台8端口交换机，其中交换机端口已经根据需要定义为属于EE VLAN或CS VLAN。但是这两台交换机应当如何互联呢？一种容易的解决方案是在每台交换机上定义一个属于CS VALN的端口（对EE VLAN也类似处理），并且如图6-26a所示将这两个端口彼此互联起来。然而，这种解决方案不具有扩展性，因为在每台交换机上N个VLAN将要求N个端口直接互联这两台交换机。



a）两条电缆



b）干线

图6-26　连接具有两个VLAN的两台VLAN交换机

一种更具扩展性互联VLAN交换机的方法称为**VLAN干线连接**（VLAN trunking）。在图6-26b所示的VLAN干线方法中，每台交换机上的一个特殊端口（左侧交换机上的端口16，右侧交换机上的端口1）被配置为干线端口，以互联这两台VLAN交换机。该干线端口属于所有VLAN，发送到任何VLAN的帧经过干线链路转发到其他交换机。但这会引起另外的问题：一个交换机怎样知道到达干线端口的帧属于某个特定的VLAN呢？IEEE定

义了一种扩展的以太网帧格式——802.1Q，用于跨越 VLAN 干线的帧。如图 6-27 中所示，802.1Q 帧由标准以太网帧与加进首部的 4 字节 **VLAN 标签**（VLAN tag）组成，而 VLAN 标签承载着该帧所属的 VLAN 标识符。VLAN 标签由在 VLAN 干线发送侧的交换机加进帧中，解析后并由在 VLAN 干线接收侧的交换机删除。VLAN 标签自身由一个 2 字节的**标签协议标识符**（Tag Protocol Identifier，TPID）字段（具有固定的十六进制值 81-00）、一个 2 字节的标签控制信息字段（包含一个 12 比特的 VLAN 标识符字段）和一个 3 比特优先权字段（具有类似于 IP 数据报 TOS 字段的目的）组成。

图 6-27　初始的以太网帧（上部），802.1Q 标签以太网 VLAN 帧（下部）

在这部分讨论中，我们仅仅简要地涉及了 VLAN，关注了基于端口的 VLAN。我们也应当提及 VLAN 能够以几种其他方式定义。在基于 MAC 的 VLAN 中，网络管理员指定属于每个 VLAN 的 MAC 地址的集合；无论何时一个设备与一个端口连接时，端口基于设备的 MAC 地址将其连接进适当的 VLAN。VLAN 也能基于网络层协议（例如 IPv4、IPv6 或 Appletalk）和其他准则进行定义。VLAN 跨越 IP 路由器扩展也是可能的，这使得多个 LAN 孤岛能被连接在一起，以形成能够跨越全局的单一 LAN［Yu 2011］。详情请参见 802.1Q 标准［IEEE 802.1q 2005］。

6.5　链路虚拟化：网络作为链路层

因为本章关注链路层协议，所以在我们临近该章结束的时候，让我们反思一下对已经演化的词汇链路的理解。在本章开始时，我们将链路视为连接两台通信主机的物理线路。在学习多路访问协议时，我们看到了多台主机能够通过一条共享的线路连接起来，并且连接主机的这种“线路”能够是无线电频谱或其他媒体。这使我们将该链路更多地抽象为一条信道，而不是作为一条线路。在我们学习以太局域网时（图 6-15），我们看到互联媒体实际上能够是一种相当复杂的交换基础设施。然而，经过这种演化，主机本身维持着这样的视图，即互联媒体只是连接两台或多台主机的链路层信道。我们看到，例如一台以太网主机不知道它是通过单一短局域网网段（图 6-17）还是通过地理上分布的交换局域网（图 6-15）或通过 VLAN 与其他局域网主机进行连接，这是很幸福的事。

在两台主机之间由拨号调制解调器连接的场合，连接这两台主机的链路实际上是电话网，这是一个逻辑上分离的、全球性的电信网络，它有自己的用于数据传输和信令的交换机、链路和协议栈。然而，从因特网链路层的观点看，通过电话网的拨号连接被看作一根简单的“线路”。在这个意义上，因特网虚拟化了电话网，将电话网看成为两台因特网主机之间提供链路层连接的链路层技术。你可能回想起在第 2 章中对于覆盖网络的讨论，类似地，一个覆盖网络将因特网视为为覆盖节点之间提供连接性的一种手段，寻求以因特网覆盖电话网的相同方式来覆盖因特网。

在本节中，我们将考虑多协议标签交换（MPLS）网络。与电路交换的电话网不同，MPLS 客观上讲是一种分组交换的虚电路网络。它们有自己的分组格式和转发行为。因此，从教学法的观点看，有关 MPLS 的讨论既适合放在网络层的学习中，也适合放在链路层的学习中。然而，从因特网的观点看，我们能够认为 MPLS 像电话网和交换以太网一样，作为为 IP 设备提供互联服务的链路层技术。因此，我们将在链路层讨论中考虑 MPLS。帧中继和 ATM 网络也能用于互联 IP 设备，虽然这些技术看上去有些过时（但仍在部署），这里将不再讨论；详情请参见一本可读性强的书［Goralski 1999］。我们对 MPLS 的讨论将是简明扼要的，因为有关这些网络每个都能够写（并且已经写了）整本书。有关 MPLS 详情我们推荐［Davie 2000］。我们这里主要关注这些网络怎样为互联 IP 设备提供服务，尽管我们也将更深入一些探讨支撑基础技术。

多协议标签交换

多协议标签交换（Multiprotocol Label Switching，MPLS）自 20 世纪 90 年代中后期在一些产业界的努力下进行演化，以改善 IP 路由器的转发速度。它采用来自虚电路网络领域的一个关键概念：固定长度标签。其目标是：对于基于固定长度标签和虚电路的技术，在不放弃基于目的地 IP 数据报转发的基础设施的前提下，当可能时通过选择性地标识数据报并允许路由器基于固定长度的标签（而不是目的地 IP 地址）转发数据报来增强其功能。重要的是，这些技术与 IP 协同工作，使用 IP 寻址和路由选择。IETF 在 MPLS 协议中统一了这些努力［RFC 3031；RFC 3032］，有效地将虚电路（VC）技术综合进了路由选择的数据报网络。

首先考虑由 MPLS 使能的路由器处理的链路层帧格式，以此开始学习 MPLS。图 6-28 显示了在 MPLS 使能的路由器之间传输的一个链路层帧，该帧具有一个小的 MPLS 首部，该首部增加到第二层（如以太网）首部和第三层（即 IP）首部之间。RFC 3032 定义了用于这种链路的 MPLS 首部的格式；用于 ATM 和帧中继网络的首部也定义在其他的 RFC 文档中。包括在 MPLS 首部中的字段是：标签；预留的 3 比特实验字段；1 比特 S 字段，用于指示一系列“成栈”的 MPLS 首部的结束（我们这里不讨论这个高级主题）；寿命字段。

图 6-28　MPLS 首部：位于链路层和网络层首部之间

从图 6-28 立即能够看出，一个 MPLS 加强的帧仅能在两个均为 MPLS 使能的路由器之间发送。（因为一个非 MPLS 使能的路由器，当它在期望发现 IP 首部的地方发现了一个 MPLS 首部时会相当混淆！）一个 MPLS 使能的路由器常被称为**标签交换路由器**（label-switched router），因为它通过在其转发表中查找 MPLS 标签，然后立即将数据报传递给适当的输出接口来转发 MPLS 帧。因此，MPLS 使能的路由器不需要提取目的 IP 地址和在转发表中执行最长前缀匹配的查找。但是路由器怎样才能知道它的邻居是否的确是 MPLS 使能的呢？路由器如何知道哪个标签与给定 IP 目的地相联系呢？为了回答这些问题，我们需要看看一组 MPLS 使能路由器之间的交互过程。

在图 6-29 所示的例子中，路由器 R1 到 R4 都是 MPLS 使能的，R5 和 R6 是标准的 IP

路由器。R1 向 R2 和 R3 通告了它（R1）能够路由到目的地 A，并且具有 MPLS 标签 6 的接收帧将要转发到目的地 A。路由器 R3 已经向路由器 R4 通告了它能够路由到目的地 A 和 D，分别具有 MPLS 标签 10 和 12 的入帧将朝着这些目的地交换。路由器 R2 也向路由器 R4 通告了它（R2）能够到达目的地 A，具有 MPLS 标签 8 的接收帧将朝着 A 交换。注意到路由器 R4 现在处于一个到达 A 且有两个 MPLS 路径的令人感兴趣的位置上，经接口 0 具有出 MPLS 标签 10，经接口 1 具有出 MPLS 标签 8。在图 6-29 中画出的外围部分是 IP 设备 R5、R6、A 和 D，它们经过一个 MPLS 基础设施（MPLS 使能路由器 R1、R2、R3 和 R4）连接在一起，这与一个交换局域网或 ATM 网络能够将 IP 设备连接到一起的方式十分相似。并且与交换局域网或 ATM 网络相似，MPLS 使能路由器 R1 到 R4 完成这些工作时从没有接触分组的 IP 首部。

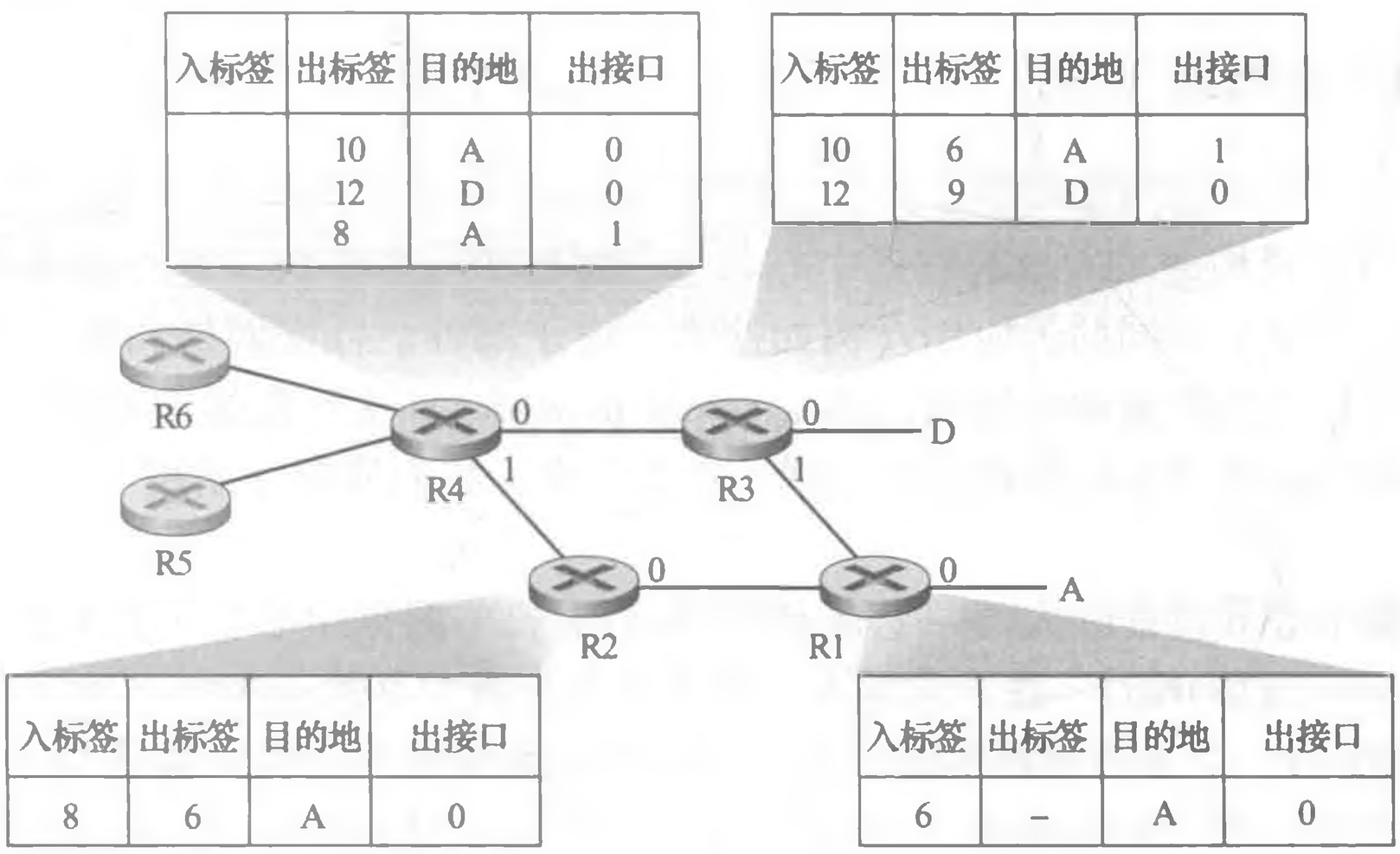


图 6-29　MPLS 增强的转发

在我们上面的讨论中，我们并没有指定在 MPLS 使能路由器之间分布标签的特定协议，因为该信令的细节已经超出了本书的范围。然而，我们注意到，IETF 的 MPLS 工作组已经在［RFC 3468］中定义了 RSVP 协议的一种扩展，称之为 RSVP-TE［RFC 3209］，它将关注对 MPLS 信令所做的工作。我们也不讨论 MPLS 实际上是如何计算在 MPLS 使能路由器之间分组的路径的，也不讨论它如何收集链路状态信息（例如，未由 MPLS 预留的链路带宽量）以用于这些路径计算中。现有的链路状态路由选择算法（例如 OSPF）已经扩展为向 MPLS 使能路由器“洪泛”。令人感兴趣的是，实际路径计算算法没有标准化，它们当前是厂商特定的算法。

至今为止，我们关于 MPLS 的讨论重点基于这样的事实，MPLS 基于标签执行交换，而不必考虑分组的 IP 地址。然而，MPLS 的真正优点和当前对 MPLS 感兴趣的原因并不在于交换速度的潜在增加，而在于 MPLS 使能的新的流量管理能力。如前面所述，R4 到 A 具有两条 MPLS 路径。如果转发在 IP 层基于 IP 地址执行，我们在第 4 章中学习的 IP 路由选择协议将只指定到 A 的单一最小费用的路径。所以，MPLS 提供了沿着多条路由转发分组的能力，使用标准 IP 路由选择协议这些路由将是不可能的。这是使用 MPLS 的一种简单形式的**流量工程**（traffic engineering）［RFC 3346；RFC3272；RFC 2702；Xiao 2000］，其中网络运行者能够超越普通的 IP 路由选择，迫使某些流量沿着一条路径朝着某给定的目的

地引导，并且朝着相同目的地的其他流量沿着另一条路径流动（无论是由于策略、性能或某些其他原因）。

将 MPLS 用于其他目的也是可能的。能用于执行 MPLS 转发路径的快速恢复，例如，经过一条预计算的无故障路径重路由流量来对链路故障做出反应［Kar 2000；Huang 2002；RFC 3469］。最后，我们注意到 MPLS 能够并且已经被用于实现所谓**虚拟专用网**（Virtual Private Network，VPN）。在为用户实现一个 VPNR 的过程中，ISP 使用它的 MPLS 使能网络将用户的各种网络连接在一起。MPLS 能被用于将资源和由用户的 VPN 使用的寻址方式相隔离，其他用户利用该 VPN 跨越该 ISP 网络，详情参见［DeClercq 2002］。

这里有关 MPLS 的讨论是简要的，我们鼓励读者查阅我们提到的这些文献。我们注意到对 MPLS 有许多可能的用途，看起来它将迅速成为因特网流量工程的瑞士军刀！

6.6　数据中心网络

近年来，因特网公司如谷歌、微软、脸书（Facebook）和亚马逊（以及它们在亚洲和欧洲的同行）已经构建了大量的数据中心。每个数据中心都容纳了数万至数十万台主机，并且同时支持着很多不同的云应用（例如搜索、电子邮件、社交网络和电子商务）。每个数据中心都有自己的**数据中心网络**（data center network），这些数据中心网络将其内部主机彼此互联并与因特网中的数据中心互联。在本节中，我们简要介绍用于云应用的数据中心网络。

大型数据中心的投资巨大，一个有 100 000 台主机的数据中心每个月的费用超过 1200 万美元［Greenberg 2009a］。在该费用中，用于主机自身的开销占 45%（每 3～4 年需要更新一次）；变压器、不间断电源系统、长时间断电时使用的发电机以及冷却系统等基础设施的开销占 25%；用于功耗的电力设施的开销占 15%；用于联网的开销占 15%，这包括了网络设备（交换机、路由器和负载均衡设备）、外部链路以及传输流量的开销。（在这些比例中，设备费用是分期偿还的，因此费用通常是由一次性购买和持续开销（如能耗）构成的。）尽管联网不是最大的费用，但是网络创新是减少整体成本和性能最大化的关键［Greenberg 2009a］。

主机就像是数据中心的工蜂：它们负责提供内容（例如，网页和视频），存储邮件和文档，并共同执行大规模分布式计算（例如，为搜索引擎提供分布式索引计算）。数据中心中的主机称为**刀片**（blade），与比萨饼盒类似，一般是包括 CPU、内存和磁盘存储的商用主机。主机被堆叠在机架上，每个机架一般堆放 20～40 台刀片。在每一个机架顶部有一台交换机，这台交换机被形象地称为**机架顶部**（Top of Rack，TOR）**交换机**，它们与机架上的主机互联，并与数据中心中的其他交换机互联。具体来说，机架上的每台主机都有一块与 TOR 交换机连接的网卡，每台 TOR 交换机有额外的端口能够与其他 TOR 交换机连接。目前主机通常用 40Gbps 的以太网连接到它们的 TOR 交换机［Greenberg 2015］。每台主机也会分配一个自己的数据中心内部的 IP 地址。

数据中心网络支持两种类型的流量：在外部客户与内部主机之间流动的流量，以及内部主机之间流动的流量。为了处理外部客户与内部主机之间流动的流量，数据中心网络包括了一台或者多台**边界路由器**（border router），它们将数据中心网络与公共因特网相连。数据中心网络因此需要将所有机架彼此互联，并将机架与边界路由器连接。图 6-30 显示了一个数据中心网络的例子。**数据中心网络设计**（data center network design）是互联网络

和协议设计的艺术，该艺术专注于机架彼此连接和与边界路由器相连。近年来，数据中心网络的设计已经成为计算机网络研究的重要分支［Al-Fares 2008；Greenberg 2009a；Greenberg 2009b；Mysore 2009；Guo 2009；Wang 2010］。

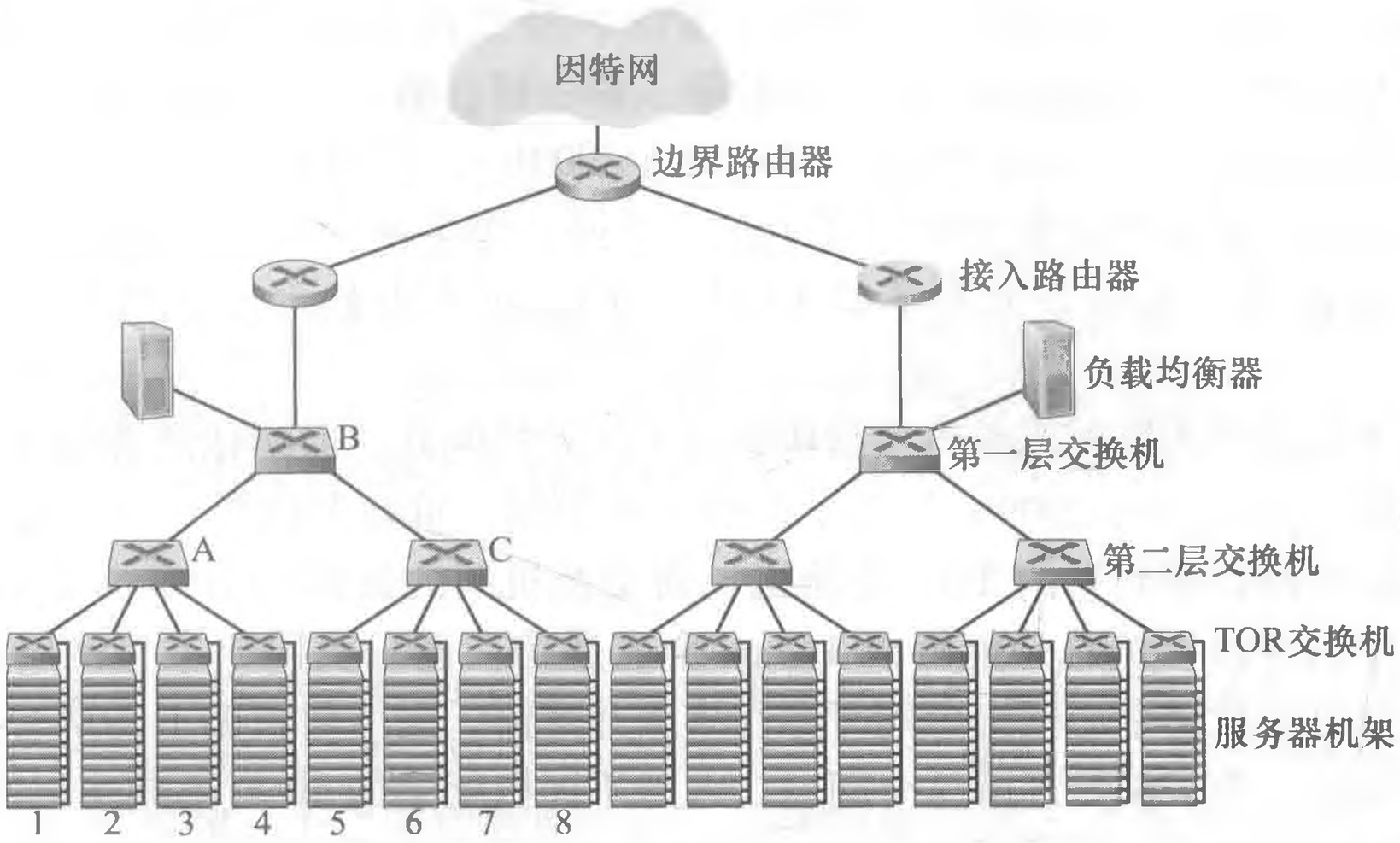


图 6-30　具有等级拓扑的数据中心网络

1. 负载均衡

一个云数据中心，如一个谷歌或者微软的数据中心，能够同时提供诸如搜索、电子邮件和视频应用等许多应用。为了支持来自外部客户的请求，每一个应用都与一个公开可见的 IP 地址关联，外部用户向该地址发送其请求并从该地址接收响应。在数据中心内部，外部请求首先被定向到一个**负载均衡器**（load balancer）。负载均衡器的任务是向主机分发请求，以主机当前的负载作为函数来在主机之间均衡负载。一个大型的数据中心通常会有几台负载均衡器，每台服务于一组特定的云应用。由于负载均衡器基于分组的目的端口号（第四层）以及目的 IP 地址做决策，因此它们常被称为“第四层交换机”。一旦接收到一个对于特定应用程序的请求，负载均衡器将该请求分发到处理该应用的某一台主机上（该主机可能再调用其他主机的服务来协助处理该请求）。当主机处理完该请求后，向负载均衡器回送响应，再由负载均衡器将其中继发回给外部客户。负载均衡器不仅平衡主机间的工作负载，而且还提供类似 NAT 的功能，将外部 IP 地址转换为内部适当主机的 IP 地址，然后将反方向流向客户的分组按照相反的转换进行处理。这防止客户直接接触主机，从而具有隐藏网络内部结构和防止客户直接与主机交互等安全性益处。

2. 等级体系结构

对于仅有数千台主机的小型数据中心，一个简单的网络也许就足够了。这种简单网络由一台边界路由器、一台负载均衡器和几十个机架组成，这些机架由单一以太网交换机进行互联。但是当主机规模扩展到几万至几十万的时候，数据中心通常应用**路由器和交换机等级结构**（hierarchy of router and switch），图 6-30 显示了这样的拓扑。在该等级结构的顶端，边界路由器与接入路由器相连（在图 6-30 中仅仅显示了两台，但是能够有更多）。在每台接入路由器下面，有 3 层交换机。每台接入路由器与一台第一层交换机相连，每台第一层交换机与多台第二层交换机以及一台负载均衡器相连。每台第二层交换机又通过机架的 TOR 交换机（第三层交换机）与多个机架相连。所有链路通常使用以太网作为链路层

和物理层协议，并混合使用铜缆和光缆。通过这种等级式设计，可以将数据中心扩展到几十万台主机的规模。

因为云应用提供商持续地提供高可用性的应用是至关重要的，所以数据中心在它们的设计中也包含了冗余网络设备和冗余链路（在图6-30中没有显示出来）。例如，每台TOR交换机能够与两台第二层交换机相连，每台接入路由器、第一层交换机和第二层交换机可以冗余并集成到设计中［Cisco 2012；Greenberg 2009b］。在图6-30中的等级设计可以看到，每台接入路由器下的这些主机构成了单一子网。为了使ARP广播流量本地化，这些子网的每个都被进一步划分为更小的VLAN子网，每个由数百台主机组成［Greenberg 2009a］。

尽管刚才描述的传统的等级体系结构解决了扩展性问题，但是依然存在主机到主机容量受限的问题［Greenberg 2009b］。为了理解这种限制，重新考虑图6-30，并且假设每台主机用1Gbps链路连接到它的TOR交换机，而交换机间的链路是10Gbps的以太网链路。在相同机架中的两台主机总是能够以1Gbps全速通信，而只受限于主机网络接口卡的速率。然而，如果在数据中心网络中同时存在多条并发流，则不同机架上的两台主机间的最大速率会小得多。为了深入理解这个问题，考虑不同机架上的40对不同主机间的40条并发流的情况。具体来说，假设图6-30中机架1上10台主机都向机架5上对应的主机发送一条流。类似地，在机架2和机架6的主机对上有10条并发流，机架3和机架7间有10条并发流，机架4和机架8间也有10条并发流。如果每一条流和其他流经同一条链路的流平均地共享链路容量，则经过10Gbps的A到B链路（以及10Gbps的B到C链路）的40条流中每条流获得的速率为10Gbps/40 = 250Mbps，显著小于1Gbps的网络接口卡速率。如果主机间的流量需要穿过该等级结构的更高层，这个问题会变得更加严重。对这个限制的一种可行的解决方案是部署更高速率的交换机和路由器。但是这会大大增加数据中心的费用，因为具有高接口速率的交换机和路由器是非常昂贵的。

因为数据中心的一个关键需求是放置计算和服务的灵活性，所以支持主机到主机的高带宽通信十分重要［Greenberg 200b；Farrington 2010］。例如，一个大规模的因特网搜索引擎可能运行在跨越多个机架的上千台主机上，在所有主机对之间具有极高的带宽要求。类似地，像EC2这样的云计算服务可能希望将构成用户服务的多台虚拟机运行在具有最大容量的物理主机上，而无须考虑它们在数据中心的位置。如果这些物理主机跨越了多个机架，前面描述的网络瓶颈可能会导致性能不佳。

3. 数据中心网络的发展趋势

为了降低数据中心的费用，同时提高其在时延和吞吐量上的性能，因特网云服务巨头如谷歌、脸书、亚马逊和微软都在不断地部署新的数据中心网络设计方案。尽管这些设计方案都是专有的，但是许多重要的趋势是一样的。

其中的一个趋势是部署能够克服传统等级设计缺陷的新型互联体系结构和网络协议。一种方法是采用**全连接拓扑**（fully connected topology）来替代交换机和路由器的等级结构［Facebook 2014；Al-Fares 2008；Greenberg 2009b；Guo 2009］，图6-31中显示了这种拓扑。在这种设计中，每台第一层交换机都与所有第二层交换机相连，因此：①主机到主机的流量绝不会超过该交换机层次；②对于n台第一层交换机，在任意两台二层交换机间有n条不相交的路径。这种设计可以显著地改善主机到主机的容量。为了理解该问题，重新考虑40条流的例子。图6-31中的拓扑能够处理这种流模式，因为在第1台第二层交换机

和第 2 台第二层交换机间存在 4 条不相交的路径，可以一起为前两台第二层交换机之间提供总和为 40Gbps 的聚合容量。这种设计不仅减轻了主机到主机的容量限制，同时创建了一种更加灵活的计算和服务环境。在这种环境中，任何未连接到同一台交换机的两个机架之间的通信在逻辑上是等价的，而不论其在数据中心的位置如何。

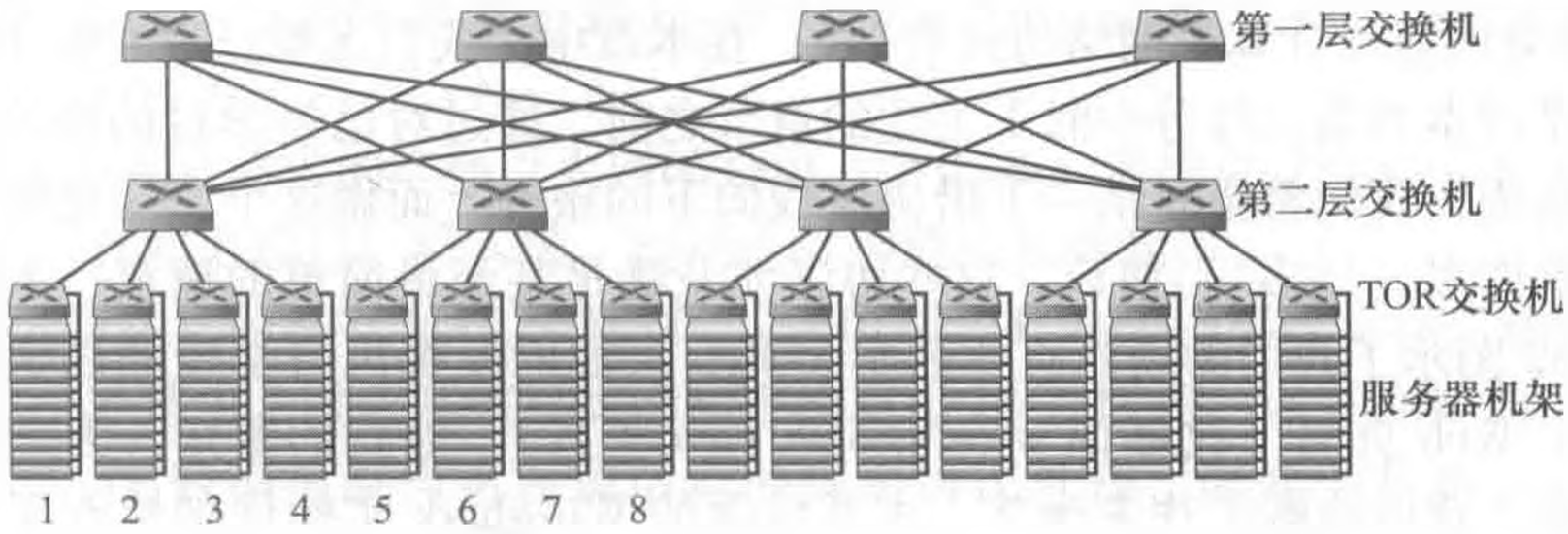


图 6-31 高度互联的数据网络拓扑

另外一个主要的趋势就是采用基于船运集装箱的**模块化数据中心**（Modular Data Center，MDC）[You Tube 2009；Waldrop 2007]。在一个 MDC 中，在一个标准的 12 米船运集装箱内，工厂构建一个“迷你数据中心”并将该集装箱运送到数据中心的位置。每一个集装箱都有多达数千台主机，堆放在数十台机架上，并且紧密地排列在一起。在数据中心位置，多个集装箱彼此互联，同时也和因特网连接。一旦预制的集装箱部署在数据中心，通常难以检修。因此，每一个集装箱都得体地设计为性能下降：当组件（服务器和交换机）随着时间的推移出现故障时，集装箱继续运行但是性能下降。当许多组件出现故障并且性能已经下降到低于某个阈值时，整个集装箱将会被移除，并用新的来替换。

创建由集装箱构成的数据中心提出了新的联网挑战。对于 MDC，有两种类型的网络：每一个集装箱中的内部网络和互联每个集装箱的核心网络［Guo 2009；Farrington 2010］。在每个集装箱内部，在规模上升到数千台主机的时候，通过廉价的商用吉比特以太网交换机创建全连接的网络（如前面所描述）是可行的。然而，核心网络的设计仍然是一个带有挑战性的问题，这需要能互联成百上千的集装箱，同时能够为典型工作负载提供跨多个集装箱的主机到主机间的高带宽。［Farrington 2010］中提出了一种互联集装箱的混合电/光交换机体系结构。

当采用高度互联拓扑的时候，一个主要的问题是设计交换机之间的路由选择算法。一种可能是采用随机路由选择方式［Greenberg 2009b］。另一种可能是在每台主机中部署多块网络接口卡［Guo 2009］，将每台主机连接到多台低成本的商用交换机上，并且允许主机自己在交换机间智能地为流量选路。这些方案的变种和扩展正被部署在当前的数据中心中。

另一种重要趋势是，大型云提供商正在其数据中心越来越多地建造或定制几乎所有东西，包括网络适配器、交换机路由器、TOR、软件和网络协议［Greenberg 2015；Singh 2015］。由亚马逊开创的另一个趋势是，用“可用性区域”来改善可靠性，这种技术在不同的邻近建筑物中基本上复制不同的数据中心。通过让建筑物邻近（几千米远），互相交互的数据能够跨越位于相同可用性区域的数据中心进行同步，与此同时提供容错性［Amazon 2014］。数据中心设计会不断出现更多的创新，感兴趣的读者可以查看近期的论文和有关数据中心设计的视频。

6.7　回顾：Web 页面请求的历程

既然我们已经在本章中学过了链路层，并且在前面几章中学过了网络层、运输层和应用层，那么我们沿协议栈向下的旅程就完成了！在本书的一开始（1.1 节），我们说过“本书的大部分内容与计算机网络协议有关”，在本章中，我们无疑已经看到了情况的确如此！在继续学习本书第二部分中时下关注的章节之前，通过对已经学过的协议做一个综合的、全面的展望，我们希望总结一下沿协议栈向下的旅程。而做这个“全面的”展望的一种方法是识别许多（许多!）协议，这些协议涉及满足甚至最简单的请求：下载一个 Web 页面。图 6-32 图示了我们的场景：一名学生 Bob 将他的便携机与学校的以太网交换机相连，下载一个 Web 页面（比如说 www. google. com 主页）。如我们所知，为满足这个看起来简单的请求，背后隐藏了许多细节。本章后面的 Wireshark 实验仔细检查了包含一些分组的踪迹文件，这些分组更为详细地涉及类似的场景。

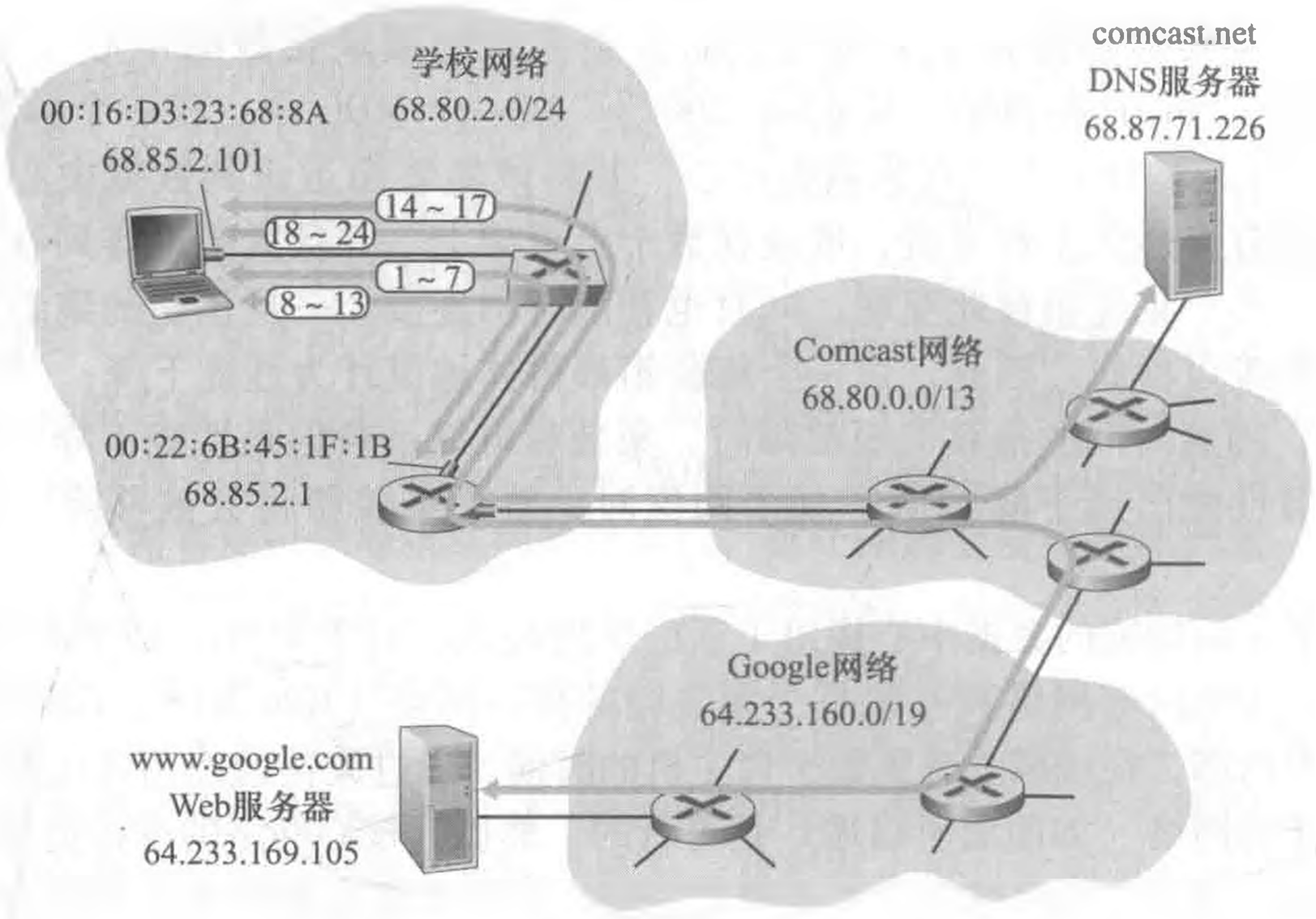


图 6-32　Web 页请求的历程：网络环境和动作

6.7.1　准备：DHCP、UDP、IP 和以太网

我们假定 Bob 启动他的便携机，然后将其用一根以太网电缆连接到学校的以太网交换机，交换机又与学校的路由器相连，如图 6-32 所示。学校的这台路由器与一个 ISP 连接，本例中 ISP 为 comcast. net。在本例中，comcast. net 为学校提供了 DNS 服务；所以，DNS 服务器驻留在 Comcast 网络中而不是学校网络中。我们将假设 DHCP 服务器运行在路由器中，就像常见情况那样。

当 Bob 首先将其便携机与网络连接时，没有 IP 地址他就不能做任何事情（例如下载一个 Web 网页）。所以，Bob 的便携机所采取的一个网络相关的动作是运行 DHCP 协议，以从本地 DHCP 服务器获得一个 IP 地址以及其他信息。

1）Bob 便携机上的操作系统生成一个 **DHCP 请求报文**（4.3.3 节），并将这个报文放入具有目的端口 67（DHCP 服务器）和源端口 68（DHCP 客户）的 **UDP 报文段**（3.3 节）

该 UDP 报文段则被放置在一个具有广播 IP 目的地址（255.255.255.255）和源 IP 地址 0.0.0.0 的 **IP 数据报**中（4.3.1 节），因为 Bob 的便携机还没有一个 IP 地址。

2）包含 DHCP 请求报文的 IP 数据报则被放置在**以太网帧**中（6.4.2 节）。该以太网帧具有目的 MAC 地址 FF:FF:FF:FF:FF:FF，使该帧将广播到与交换机连接的所有设备（如果顺利的话也包括 DHCP 服务器）；该帧的源 MAC 地址是 Bob 便携机的 MAC 地址 00:16:D3:23:68:8A。

3）包含 DHCP 请求的广播以太网帧是第一个由 Bob 便携机发送到以太网交换机的帧。该交换机在所有的出端口广播入帧，包括连接到路由器的端口。

4）路由器在它的具有 MAC 地址 00:22:6B:45:1F 的接口接收到该广播以太网帧，该帧中包含 DHCP 请求，并且从该以太网帧中抽取出 IP 数据报。该数据报的广播 IP 目的地址指示了这个 IP 数据报应当由在该节点的高层协议处理，因此该数据报的载荷（一个 UDP 报文段）**被分解**（3.2 节）向上到达 UDP，DHCP 请求报文从此 UDP 报文段中抽取出来。此时 DHCP 服务器有了 DHCP 请求报文。

5）我们假设运行在路由器中的 DHCP 服务器能够以 CIDR（4.3.3 节）块 68.85.2.0/24 分配 IP 地址。所以本例中，在学校内使用的所有 IP 地址都在 Comcast 的地址块中。我们假设 DHCP 服务器分配地址 68.85.2.101 给 Bob 的便携机。DHCP 服务器生成包含这个 IP 地址以及 DNS 服务器的 IP 地址（68.87.71.226）、默认网关路由器的 IP 地址（68.85.2.1）和子网块（68.85.2.0/24）（等价为“网络掩码”）的一个 DHCP ACK **报文**（4.3.3 节）。该 DHCP 报文被放入一个 UDP 报文段中，UDP 报文段被放入一个 IP 数据报中，IP 数据报再被放入一个以太网帧中。这个以太网帧的源 MAC 地址是路由器连到归属网络时接口的 MAC 地址（00:22:6B:45:1F:1B），目的 MAC 地址是 Bob 便携机的 MAC 地址（00:16:D3:23:68:8A）。

6）包含 DHCP ACK 的以太网帧由路由器发送给交换机。因为交换机是**自学习**的（6.4.3 节），并且先前从 Bob 便携机收到（包含 DHCP 请求的）以太网帧，所以该交换机知道寻址到 00:16:D3:23:68:8A 的帧仅从通向 Bob 便携机的输出端口转发。

7）Bob 便携机接收到包含 DHCP ACK 的以太网帧，从该以太网帧中抽取 IP 数据报，从 IP 数据报中抽取 UDP 报文段，从 UDP 报文段抽取 DHCP ACK 报文。Bob 的 DHCP 客户则记录下它的 IP 地址和它的 DNS 服务器的 IP 地址。它还在其 **IP 转发表**中安装默认网关的地址（4.1 节）。Bob 便携机将向该默认网关发送目的地址为其子网 68.85.2.0/24 以外的所有数据报。此时，Bob 便携机已经初始化好它的网络组件，并准备开始处理 Web 网页获取。（注意到在第 4 章给出的四个步骤中仅有最后两个 DHCP 步骤是实际必要的。）

6.7.2 仍在准备：DNS 和 ARP

当 Bob 将 www.google.com 的 URL 键入其 Web 浏览器时，他开启了一长串事件，这将导致谷歌主页最终显示在其 Web 浏览器上。Bob 的 Web 浏览器通过生成一个 **TCP 套接字**（2.7 节）开始了该过程，套接字用于向 www.google.com 发送 **HTTP 请求**（2.2 节）。为了生成该套接字，Bob 便携机将需要知道 www.google.com 的 IP 地址。我们在 2.4 节中学过，使用 **DNS 协议**提供这种名字到 IP 地址的转换服务。

8）Bob 便携机上的操作系统因此生成一个 **DNS 查询报文**（2.4.3 节），将字符串 www.google.com 放入 DNS 报文的问题段中。该 DNS 报文则放置在一个具有 53 号（DNS 服务器）目的端口的 UDP 报文段中。该 UDP 报文段则被放入具有 IP 目的地址 68.87.71.226（在