（而不是几十米）范围内。

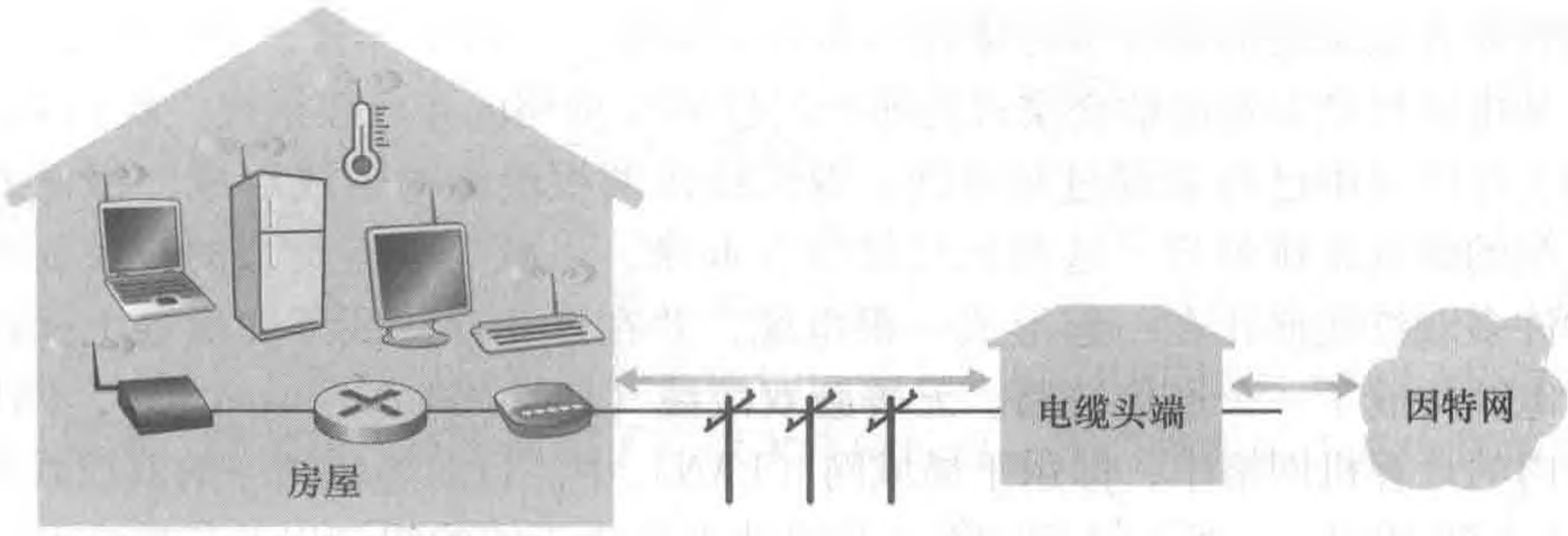

图 1-9　一个典型的家庭网络的示意图

电信公司已经在所谓第三代（3G）无线技术中进行了大量投资，3G 为分组交换广域无线因特网接入提供了超过 1Mbps 的速率。甚至更高速率的广域接入技术即第四代（4G）广域无线网络也已经被部署了。LTE（长期演进“Long-Term Evolution”的缩写，被评为最差首字母缩写词年度奖候选者）来源于 3G 技术，它能够取得超过 10Mbps 的速率。据报道，几十 Mbps 的 LTE 下行速率已经在商业部署中得到应用。我们将在第 7 章讨论无线网络和移动性，以及 WiFi、3G 和 LTE 等技术的基本原则。

1.2.2　物理媒体

在前面的内容中，我们概述了因特网中某些最为重要的网络接入技术。当我们描述这些技术时，我们也指出了所使用的物理媒体。例如，我们说过 HFC 使用了光缆和同轴电缆相结合的技术。我们说过 DSL 和以太网使用了双绞铜线。我们也说过移动接入网使用了无线电频谱。在这一节中，我们简要概述一下这些和其他常在因特网中使用的传输媒体。

为了定义物理媒体所表示的内容，我们仔细思考一下一个比特的短暂历程。考虑一个比特从一个端系统开始传输，通过一系列链路和路由器，到达另一个端系统。这个比特被漫不经心地传输了许许多多次！源端系统首先发射这个比特，不久后其中的第一台路由器接收该比特；第一台路由器发射该比特，接着不久后第二台路由器接收该比特；等等。因此，这个比特当从源到目的地传输时，通过一系列“发射器 - 接收器”对。对于每个发射器 - 接收器对，通过跨越一种**物理媒体**（physical medium）传播电磁波或光脉冲来发送该比特。该物理媒体可具有多种形状和形式，并且对沿途的每个发射器 - 接收器对而言不必具有相同的类型。物理媒体的例子包括双绞铜线、同轴电缆、多模光纤缆、陆地无线电频谱和卫星无线电频谱。物理媒体分成两种类型：**导引型媒体**（guided media）和**非导引型媒体**（unguided media）。对于导引型媒体，电波沿着固体媒体前行，如光缆、双绞铜线或同轴电缆。对于非导引型媒体，电波在空气或外层空间中传播，例如在无线局域网或数字卫星频道中。

在深入讨论各种媒体类型的特性之前，我们简要地讨论一下它们的成本。物理链路（铜线、光缆等）的实际成本与其他网络成本相比通常是相当小的。特别是安装物理链路的劳动力成本能够比材料的成本高出几个数量级。正因为这个原因，许多建筑商在一个建筑物的每个房间中安装了双绞线、光缆和同轴电缆。即使最初仅使用了一种媒体，在不久的将来也可能会使用另一种媒体，这样将来不必再铺设另外的线缆，从而节省了经费。

1. 双绞铜线

最便宜并且最常用的导引型传输媒体是双绞铜线。一百多年来，它一直用于电话网。事实上，从电话机到本地电话交换机的连线超过99%使用的是双绞铜线。我们多数人在自己家中和工作环境中已经看到过双绞线。双绞线由两根绝缘的铜线组成，每根大约1mm粗，以规则的螺旋状排列着。这两根线被绞合起来，以减少邻近类似的双绞线的电气干扰。通常许多双绞线捆扎在一起形成一根电缆，并在这些双绞线外面覆盖上保护性防护层。一对电线构成了一个通信链路。**无屏蔽双绞线**（Unshielded Twisted Pair，UTP）常用在建筑物内的计算机网络中，即用于局域网（LAN）中。目前局域网中的双绞线的数据速率从10Mbps到10Gbps。所能达到的数据传输速率取决于线的粗细以及传输方和接收方之间的距离。

20世纪80年代出现光纤技术时，许多人因为双绞线比特速率低而轻视它，某些人甚至认为光纤技术将完全代替双绞线。但双绞线不是那么容易被抛弃的。现代的双绞线技术例如6a类电缆能够达到10Gbps的数据传输速率，距离长达100m。双绞线最终已经作为高速LAN联网的主导性解决方案。

如前面讨论的那样，双绞线也经常用于住宅因特网接入。我们看到，拨号调制解调器技术通过双绞线能以高达56kbps的速率接入。我们也看到，数字用户线（DSL）技术通过双绞线使住宅用户以超过数十Mbps的速率接入因特网（当用户靠近ISP的中心局居住时）。

2. 同轴电缆

与双绞线类似，同轴电缆由两个铜导体组成，但是这两个导体是同心的而不是并行的。借助于这种结构及特殊的绝缘体和保护层，同轴电缆能够达到较高的数据传输速率。同轴电缆在电缆电视系统中相当普遍。我们前面已经看到，电缆电视系统最近与电缆调制解调器结合起来，为住宅用户提供数十Mbps速率的因特网接入。在电缆电视和电缆因特网接入中，发送设备将数字信号调制到某个特定的频段，产生的模拟信号从发送设备传送到一个或多个接收方。同轴电缆能被用作导引型**共享媒体**（shared medium）。特别是，许多端系统能够直接与该电缆相连，每个端系统都能接收由其他端系统发送的内容。

3. 光纤

光纤是一种细而柔软的、能够导引光脉冲的媒体，每个脉冲表示一个比特。一根光纤能够支持极高的比特速率，高达数十甚至数百Gbps。它们不受电磁干扰，长达100km的光缆信号衰减极低，并且很难窃听。这些特征使得光纤成为长途导引型传输媒体，特别是跨海链路。在美国和别的地方，许多长途电话网络现在全面使用光纤。光纤也广泛用于因特网的主干。然而，高成本的光设备，如发射器、接收器和交换机，阻碍光纤在短途传输中的应用，如在LAN或家庭接入网中就不使用它们。光载波（Optical Carrier，OC）标准链路速率的范围从51.8Mbps到39.8Gbps；这些标准常被称为OC-n，其中的链路速率等于$n \times 51.8$Mbps。目前正在使用的标准包括OC-1、OC-3、OC-12、OC-24、OC-48、OC-96、OC-192、OC-768。[Mukherjee 2006；Ramaswami 2010]提供了光纤网络各方面的知识。

4. 陆地无线电信道

无线电信道承载电磁频谱中的信号。它不需要安装物理线路，并具有穿透墙壁、提供

与移动用户的连接以及长距离承载信号的能力，因而成为一种有吸引力的媒体。无线电信道的特性极大地依赖于传播环境和信号传输的距离。环境上的考虑取决于路径损耗和遮挡衰落（即当信号跨距离传播和绕过/通过阻碍物体时信号强度降低）、多径衰落（由于干扰对象的信号反射）以及干扰（由于其他传输或电磁信号）。

陆地无线电信道能够大致划分为三类：一类运行在很短距离（如 1 米或 2 米）；另一类运行在局域，通常跨越数十到几百米；第三类运行在广域，跨越数万米。个人设备如无线头戴式耳机、键盘和医疗设备跨短距离运行；在 1.2.1 节中描述的无线 LAN 技术使用了局域无线电信道；蜂窝接入技术使用了广域无线电信道。我们将在第 7 章中详细讨论无线电信道。

5. 卫星无线电信道

一颗通信卫星连接地球上的两个或多个微波发射器/接收器，它们被称为地面站。该卫星在一个频段上接收传输，使用一个转发器（下面讨论）再生信号，并在另一个频率上发射信号。通信中常使用两类卫星：**同步卫星**（geostationary satellite）和**近地轨道**（Low-Earth Orbiting，LEO）卫星［Wiki Satellite 2016］。

同步卫星永久地停留在地球上方的相同点上。这种静止性是通过将卫星置于地球表面上方 36 000km 的轨道上而取得的。从地面站到卫星再回到地面站的巨大距离引入了可观的 280ms 信号传播时延。不过，能以数百 Mbps 速率运行的卫星链路通常用于那些无法使用 DSL 或电缆因特网接入的区域。

近地轨道卫星放置得非常靠近地球，并且不是永久地停留在地球上方的一个点。它们围绕地球旋转，就像月亮围绕地球旋转那样，并且彼此之间可进行通信，也可以与地面站通信。为了提供对一个区域的连续覆盖，需要在轨道上放置许多卫星。当前有许多低轨道通信系统在研制中。LEO 卫星技术未来也许能够用于因特网接入。

1.3 网络核心

在考察了因特网边缘后，我们现在更深入地研究网络核心，即由互联因特网端系统的分组交换机和链路构成的网状网络。图 1-10 用加粗阴影线勾画出网络核心部分。

1.3.1 分组交换

在各种网络应用中，端系统彼此交换**报文**（message）。报文能够包含协议设计者需要的任何东西。报文可以执行一种控制功能（例如，图 1-2 所示例子中的“你好”报文），也可以包含数据，例如电子邮件数据、JPEG 图像或 MP3 音频文件。为了从源端系统向目的端系统发送一个报文，源将长报文划分为较小的数据块，称之为**分组**（packet）。在源和目的地之间，每个分组都通过通信链路和**分组交换机**（packet switch）传送。（交换机主要有两类：**路由器**（router）和**链路层交换机**（link-layer switch）。）分组以等于该链路最大传输速率的速度传输通过通信链路。因此，如果某源端系统或分组交换机经过一条链路发送一个 L 比特的分组，链路的传输速率为 R 比特/秒，则传输该分组的时间为 L/R 秒。

1. 存储转发传输

多数分组交换机在链路的输入端使用**存储转发传输**（store-and-forward transmission）机制。存储转发传输是指在交换机能够开始向输出链路传输该分组的第一个比特之前，必

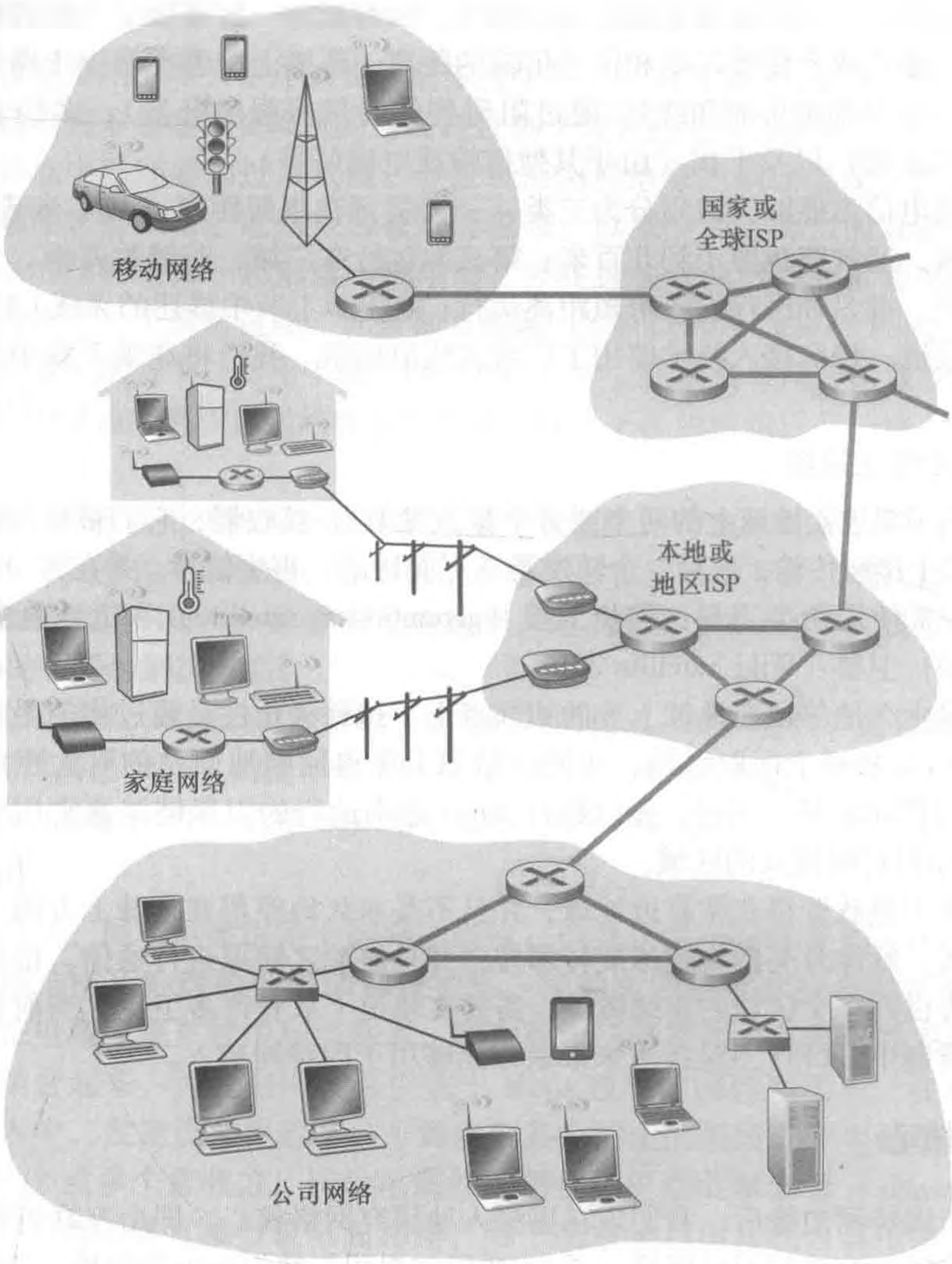

图 1-10　网络核心

须接收到整个分组。为了更为详细地探讨存储转发传输，考虑由两个端系统经一台路由器连接构成的简单网络，如图 1-11 所示。一台路由器通常有多条繁忙的链路，因为它的任务就是把一个入分组交换到一条出链路。在这个简单例子中，该路由器的任务相当简单：将分组从一条（输入）链路转移到另一条唯一的连接链路。在图 1-11 所示的特定时刻，源已经传输了分组 1 的一部分，分组 1 的前沿已经到达了路由器。因为该路由器应用了存储转发机制，所以此时它还不能传输已经接收的比特，而是必须先缓存（即“存储”）该分组的比特。仅当路由器已经接收完了该分组的所有比特后，它才能开始向出链路传输（即“转发”）该分组。为了深刻领悟存储转发传输，我们现在计算一下从源开始发送分组到目的地收到整个分组所经过的时间。（这里我们将忽略传播时延——指这些比特以接近光速的

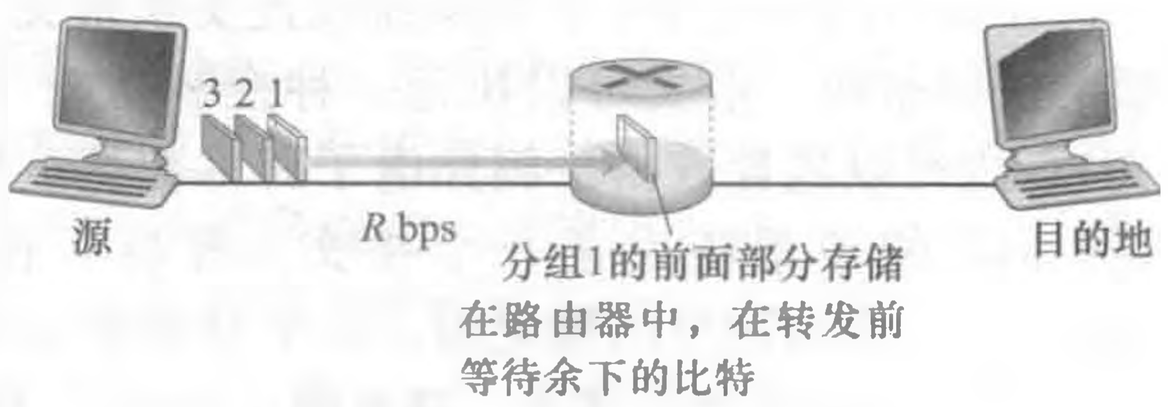

图 1-11　存储转发分组交换

速度跨越线路所需要的时间，这将在 1.4 节讨论。）源在时刻 0 开始传输，在时刻 L/R 秒，因为该路由器刚好接收到整个分组，所以它能够朝着目的地向出链路开始传输分组；在时刻 $2L/R$，路由器已经传输了整个分组，并且整个分组已经被目的地接收。所以，总时延是 $2L/R$。如果交换机一旦比特到达就转发比特（不必首先收到整个分组），则因为比特没有在路由器保存，总时延将是 L/R。但是如我们将在 1.4 节中讨论的那样，路由器在转发前需要接收、存储和处理整个分组。

现在我们来计算从源开始发送第一个分组到目的地接收到所有三个分组所需的时间。与前面一样，在时刻 L/R，路由器开始转发第一个分组。而在时刻 L/R 源也开始发送第二个分组，因为它已经完成了第一个分组的完整发送。因此，在时刻 $2L/R$，目的地已经收到第一个分组并且路由器已经收到第二个分组。类似地，在时刻 $3L/R$，目的地已经收到前两个分组并且路由器已经收到第三个分组。最后，在时刻 $4L/R$，目的地已经收到所有 3 个分组！

我们现在来考虑下列一般情况：通过由 N 条速率均为 R 的链路组成的路径（所以，在源和目的地之间有 $N-1$ 台路由器），从源到目的地发送一个分组。应用如上相同的逻辑，我们看到端到端时延是：

$$d_{\text{端到端}} = N\frac{L}{R} \tag{1-1}$$

你也许现在要试着确定 P 个分组经过 N 条链路序列的时延有多大。

2. 排队时延和分组丢失

每台分组交换机有多条链路与之相连。对于每条相连的链路，该分组交换机具有一个**输出缓存**（output buffer，也称为**输出队列**（output queue）），它用于存储路由器准备发往那条链路的分组。该输出缓存在分组交换中起着重要的作用。如果到达的分组需要传输到某条链路，但发现该链路正忙于传输其他分组，该到达分组必须在输出缓存中等待。因此，除了存储转发时延以外，分组还要承受输出缓存的**排队时延**（queuing delay）。这些时延是变化的，变化的程度取决于网络的拥塞程度。因为缓存空间的大小是有限的，一个到达的分组可能发现该缓存已被其他等待传输的分组完全充满了。在此情况下，将出现**分组丢失（丢包）**（packet loss），到达的分组或已经排队的分组之一将被丢弃。

图 1-12 显示了一个简单的分组交换网络。如在图 1-11 中，分组被表示为三维厚片。厚片的宽度表示了该分组中比特的数量。在这张图中，所有分组具有相同的宽度，因此有相同的长度。假定主机 A 和 B 向主机 E 发送分组。主机 A 和 B 先通过 100Mbps 的以太网链路向第一个路由器发送分组。该路由器则将这些分组导向到一条 15Mbps 的链路。在某个短时间间隔内，如果分组到达路由器的到达率（转换为每秒比特）超过了 15Mbps，这些分组在通过链路传输之前，将在链路输出缓存中排队，在该路由器中将出现拥塞。例如，如果主机 A 和主机 B 每个都同时发送了 5 个紧接着的分组突发块，则这些分组中的大多数将在队列中等待一些时间。事实上，这完全类似于每天都在经历的一些情况，例如当我们在银行柜台前排队等待或在过路收费站前等待时。我们将在 1.4 节中更为详细地研究这种排队时延。

3. 转发表和路由选择协议

前面我们说过，路由器从与它相连的一条通信链路得到分组，然后向与它相连的另一条通信链路转发该分组。但是路由器怎样决定它应当向哪条链路进行转发呢？不同的计算

机网络实际上是以不同的方式完成分组转发的。这里，我们简要介绍在因特网中所采用的方法。

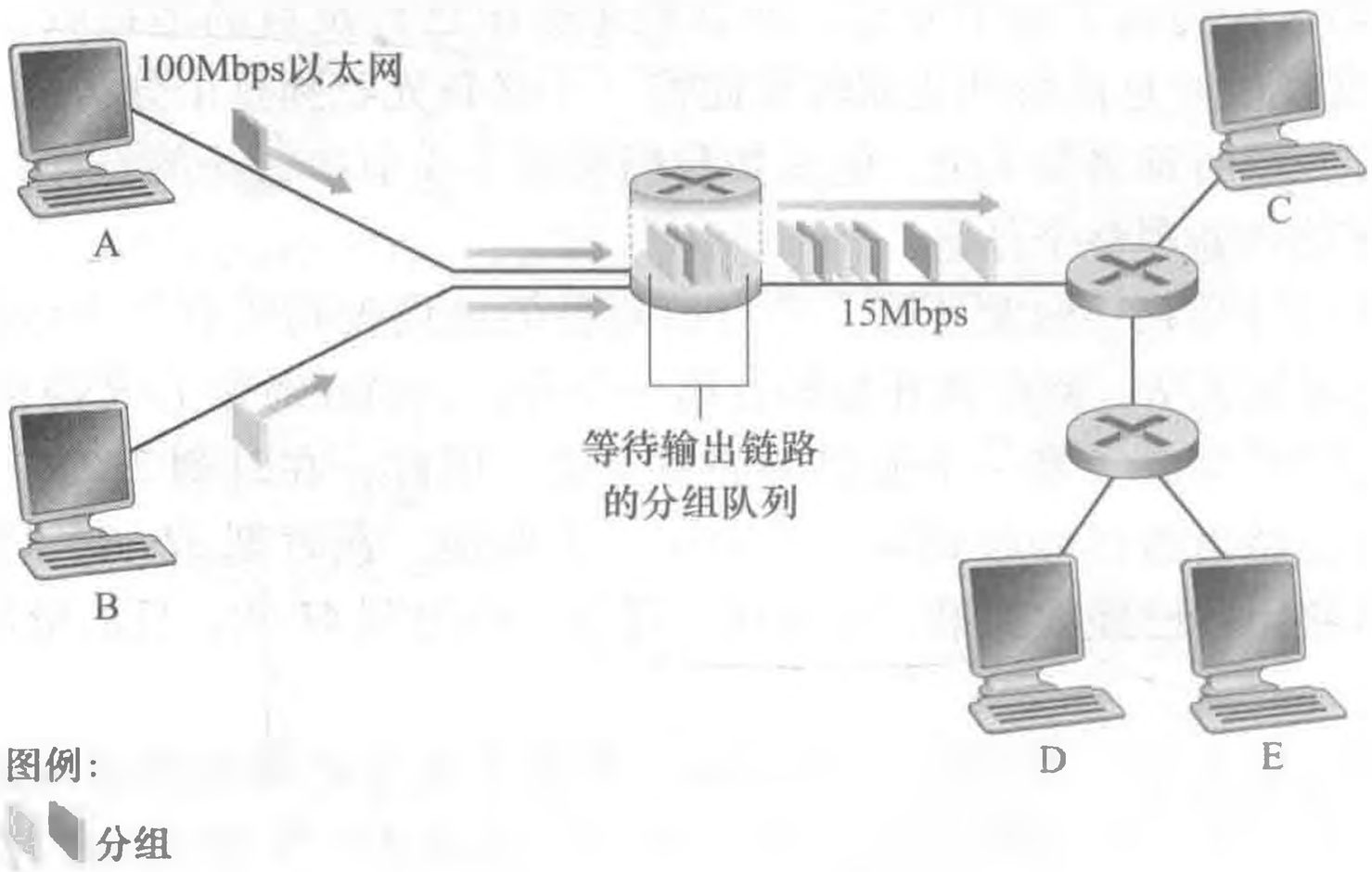

图 1-12　分组交换

在因特网中，每个端系统具有一个称为 IP 地址的地址。当源主机要向目的端系统发送一个分组时，源在该分组的首部包含了目的地的 IP 地址。如同邮政地址那样，该地址具有一种等级结构。当一个分组到达网络中的路由器时，路由器检查该分组的目的地址的一部分，并向一台相邻路由器转发该分组。更特别的是，每台路由器具有一个**转发表**（forwarding table），用于将目的地址（或目的地址的一部分）映射成为输出链路。当某分组到达一台路由器时，路由器检查该地址，并用这个目的地址搜索其转发表，以发现适当的出链路。路由器则将分组导向该出链路。

端到端选路过程可以用一个不使用地图而喜欢问路的汽车驾驶员来类比。例如，假定 Joe 驾车从费城到佛罗里达州奥兰多市的 Lakeside Drive 街 156 号。Joe 先驾车到附近的加油站，询问怎样才能到达佛罗里达州奥兰多市的 Lakeside Drive 街 156 号。加油站的服务员从该地址中抽取了佛罗里达州部分，告诉 Joe 他需要上 I-95 南州际公路，该公路恰有一个邻近该加油站的入口。他又告诉 Joe，一旦到了佛罗里达后应当再问当地人。于是，Joe 上了 I-95 南州际公路，一直到达佛罗里达的 Jacksonville，在那里他向另一个加油站服务员问路。该服务员从地址中抽取了奥兰多市部分，告诉 Joe 他应当继续沿 I-95 公路到 Daytona 海滩，然后再问其他人。在 Daytona 海滩，另一个加油站服务员也抽取该地址的奥兰多部分，告诉 Joe 应当走 I-4 公路直接前往奥兰多。Joe 走了 I-4 公路，并从奥兰多出口下来。Joe 又向另一个加油站的服务员询问，这时该服务员抽取了该地址的 Lakeside Drive 部分，告诉了 Joe 到 Lakeside Drive 必须要走的路。一旦 Joe 到达了 Lakeside Drive，他向一个骑自行车的小孩询问了到达目的地的方法。这个孩子抽取了该地址的 156 号部分，并指示了房屋的方向。Joe 最后到达了最终目的地。在上述类比中，那些加油站服务员和骑车的孩子可类比为路由器。

我们刚刚学习了路由器使用分组的目的地址来索引转发表并决定适当的出链路。但是这个叙述还要求回答另一个问题：转发表是如何进行设置的？是通过人工对每台路由器逐台进行配置，还是因特网使用更为自动的过程进行配置呢？第 5 章将深入探讨这个问题。但在这里为了激发你的求知欲，我们现在将告诉你因特网具有一些特殊的**路由选择协议**

(routing protocol)，用于自动地设置这些转发表。例如，一个路由选择协议可以决定从每台路由器到每个目的地的最短路径，并使用这些最短路径结果来配置路由器中的转发表。

怎样才能实际看到分组在因特网中所走的端到端路由呢？我们现在请你亲手用一下 Traceroute 程序。直接访问站点 www. traceroute. org，在一个特定的国家中选择一个源，跟踪从该源到你的计算机的路由。(参见 1.4 节有关 Traceroute 的讨论。)

1.3.2　电路交换

通过网络链路和交换机移动数据有两种基本方法：**电路交换** (circuit switching) 和**分组交换** (packet switching)。上一小节已经讨论过分组交换网络，现在我们将注意力转向电路交换网络。

在电路交换网络中，在端系统间通信会话期间，预留了端系统间沿路径通信所需要的资源（缓存，链路传输速率)。在分组交换网络中，这些资源则不是预留的；会话的报文按需使用这些资源，其后果可能是不得不等待（即排队）接入通信线路。一个简单的类比是，考虑两家餐馆，一家需要顾客预订，而另一家不需要预订，但不保证能安排顾客。对于需要预订的那家餐馆，我们在离开家之前必须承受先打电话预订的麻烦，但当我们到达该餐馆时，原则上我们能够立即入座并点菜。对于不需要预订的那家餐馆，我们不必麻烦地预订餐桌，但当我们到达该餐馆时，也许不得不先等待一张餐桌空闲后才能入座。

传统的电话网络是电路交换网络的例子。考虑当一个人通过电话网向另一个人发送信息（语音或传真）时所发生的情况。在发送方能够发送信息之前，该网络必须在发送方和接收方之间建立一条连接。这是一个名副其实的连接，因为此时沿着发送方和接收方之间路径上的交换机都将为该连接维护连接状态。用电话的术语来说，该连接被称为一条**电路** (circuit)。当网络创建这种电路时，它也在连接期间在该网络链路上预留了恒定的传输速率（表示为每条链路传输容量的一部分)。既然已经为该发送方 - 接收方连接预留了带宽，则发送方能够以确保的恒定速率向接收方传送数据。

图 1-13 显示了一个电路交换网络。在这个网络中，用 4 条链路互联了 4 台电路交换机。这些链路中的每条都有 4 条电路，因此每条链路能够支持 4 条并行的连接。每台主机（例如 PC 和工作站）都与一台交换机直接相连。当两台主机要通信时，该网络在两台主机之间创建一条专用的**端到端连接** (end- to- end connection)。因此，主机 A 为了向主机 B 发送报文，网络必须在两条链路的每条上先预留一条电路。在这个例子中，这条专用的端到端连接使用第一条链路中的第二条电路和第二条链路中的第四条电路。因为每条链路具有 4 条电路，对于由端到端连接所使用的每条链路而言，该连接在连接期间获得链路总传输容量的 1/4。例如，如果两台邻近交换机之间每条链路具有 1Mbps 传输速率，则每个端到端电路交换连接获得 250kbps 专用的传输速率。

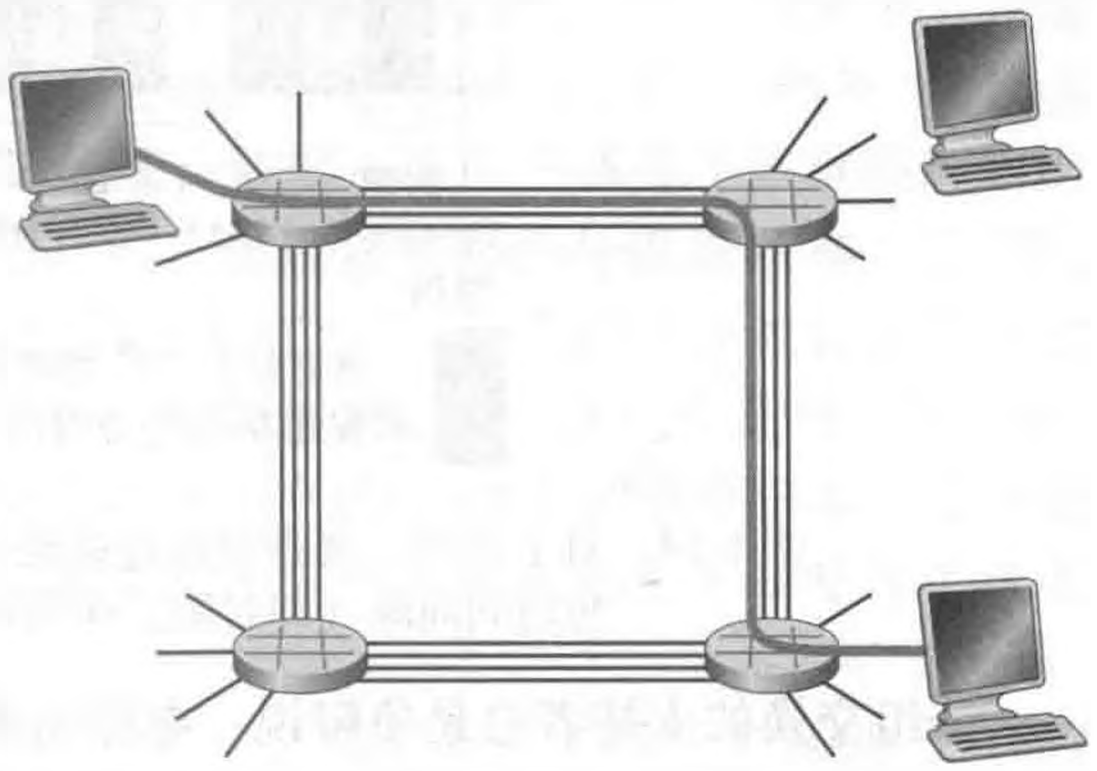

图 1-13　由 4 台交换机和 4 条链路组成的一个简单电路交换网络

与此相反，考虑一台主机要经过分组交换网络（如因特网）向另一台主机发送分组所发生的情况。与使用电路交换相同，该分组经过一系列通信链路传输。但与电路交换不同

的是，该分组被发送进网络，而不预留任何链路资源之类的东西。如果因为此时其他分组也需要经该链路进行传输而使链路之一出现拥塞，则该分组将不得不在传输链路发送侧的缓存中等待而产生时延。因特网尽最大努力以实时方式交付分组，但它不做任何保证。

1. 电路交换网络中的复用

链路中的电路是通过**频分复用**（Frequency-Division Multiplexing，FDM）或**时分复用**（Time-Division Multiplexing，TDM）来实现的。对于 FDM，链路的频谱由跨越链路创建的所有连接共享。特别是，在连接期间链路为每条连接专用一个频段。在电话网络中，这个频段的宽度通常为 4kHz（即每秒 4000 周期）。毫无疑问，该频段的宽度称为**带宽**（bandwidth）。调频无线电台也使用 FDM 来共享 88MHz～108MHz 的频谱，其中每个电台被分配一个特定的频段。

对于一条 TDM 链路，时间被划分为固定期间的帧，并且每个帧又被划分为固定数量的时隙。当网络跨越一条链路创建一条连接时，网络在每个帧中为该连接指定一个时隙。这些时隙专门由该连接单独使用，一个时隙（在每个帧内）可用于传输该连接的数据。

图 1-14 显示了一个支持多达 4 条电路的特定网络链路的 FDM 和 TDM。对于 FDM，其频率域被分割为 4 个频段，每个频段的带宽是 4kHz。对于 TDM，其时域被分割为帧，在每个帧中具有 4 个时隙，在循环的 TDM 帧中每条电路被分配相同的专用时隙。对于 TDM，一条电路的传输速率等于帧速率乘以一个时隙中的比特数量。例如，如果链路每秒传输 8000 个帧，每个时隙由 8 个比特组成，则每条电路的传输速率是 64kbps。

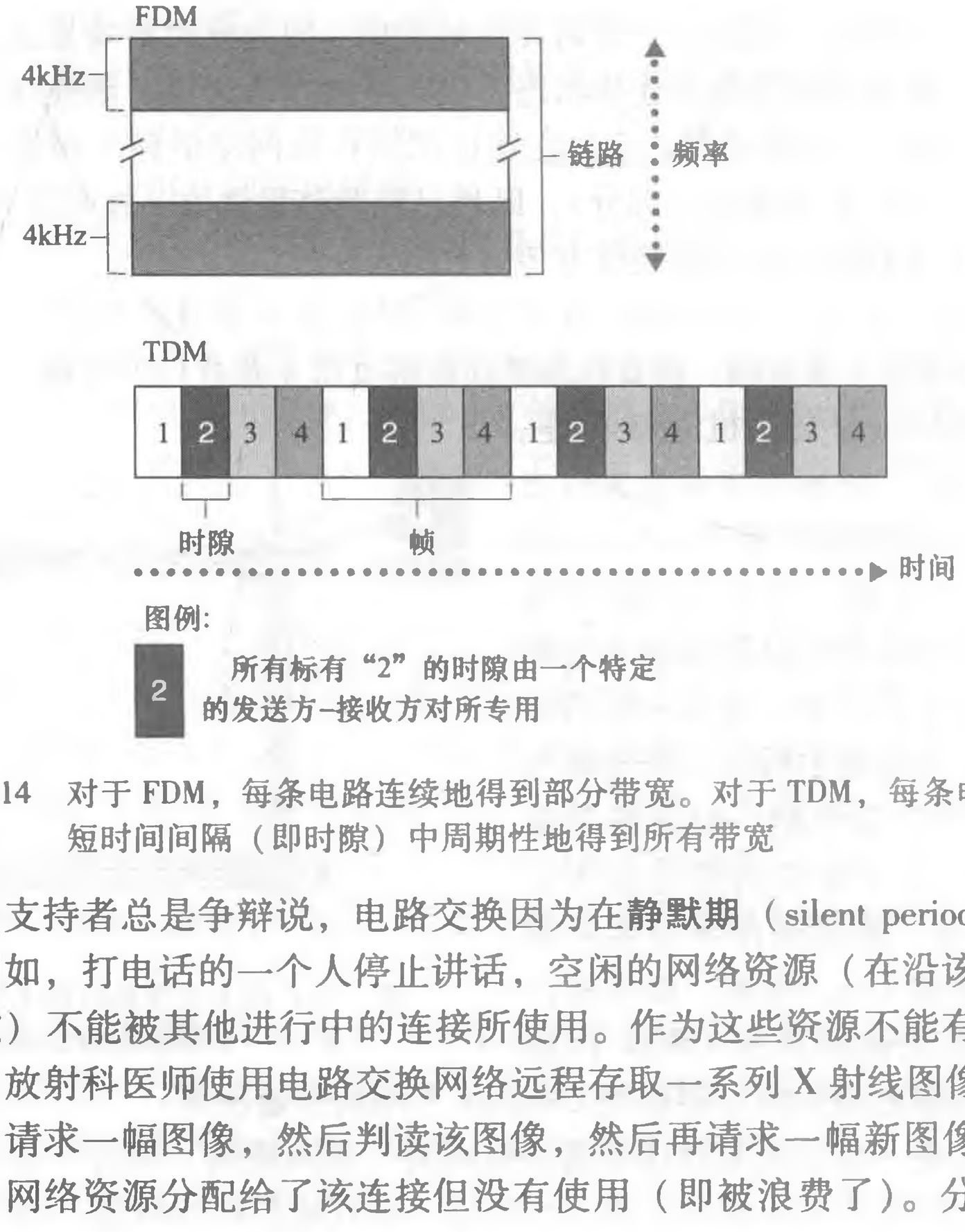

图 1-14　对于 FDM，每条电路连续地得到部分带宽。对于 TDM，每条电路在短时间间隔（即时隙）中周期性地得到所有带宽

分组交换的支持者总是争辩说，电路交换因为在**静默期**（silent period）专用电路空闲而不够经济。例如，打电话的一个人停止讲话，空闲的网络资源（在沿该连接路由的链路中的频段或时隙）不能被其他进行中的连接所使用。作为这些资源不能有效利用的另一个例子，考虑一名放射科医师使用电路交换网络远程存取一系列 X 射线图像。该放射科医师建立一条连接，请求一幅图像，然后判读该图像，然后再请求一幅新图像。在放射科医师判读图像期间，网络资源分配给了该连接但没有使用（即被浪费了）。分组交换的支持者

还津津乐道地指出，创建端到端电路和预留端到端带宽是复杂的，需要复杂的信令软件以协调沿端到端路径的交换机的操作。

在结束讨论电路交换之前，我们讨论一个用数字表示的例子，它更能说明问题的实质。考虑从主机 A 到主机 B 经一个电路交换网络发送一个 640 000 比特的文件需要多长时间。假如在该网络中所有链路使用具有 24 时隙的 TDM，比特速率为 1.536Mbps。同时假定在主机 A 能够开始传输该文件之前，需要 500ms 创建一条端到端电路。它需要多长时间才能发送该文件？每条链路具有的传输速率是 1.536Mbps/24 = 64kbps，因此传输该文件需要 (640kb)/(64kbps) = 10s。这个 10s，再加上电路创建时间，这样就需要 10.5s 发送该文件。值得注意的是，该传输时间与链路数量无关：端到端电路不管是通过一条链路还是 100 条链路，传输时间都将是 10s。(实际的端到端时延还包括传播时延，参见 1.4 节。)

2. 分组交换与电路交换的对比

在描述了电路交换和分组交换之后，我们来对比一下这两者。分组交换的批评者经常争辩说，分组交换不适合实时服务（例如，电话和视频会议），因为它的端到端时延是可变的和不可预测的（主要是因为排队时延的变动和不可预测所致）。分组交换的支持者却争辩道：①它提供了比电路交换更好的带宽共享；②它比电路交换更简单、更有效，实现成本更低。分组交换与电路交换之争的有趣讨论参见 [Molinero-Fernandez 2002]。概括而言，嫌餐馆预订麻烦的人宁可要分组交换而不愿意要电路交换。

分组交换为什么更有效呢？我们看一个简单的例子。假定多个用户共享一条 1Mbps 链路，再假定每个用户活跃周期是变化的，某用户时而以 100kbps 恒定速率产生数据，时而静止——这时用户不产生数据。进一步假定该用户仅有 10% 的时间活跃（余下的 90% 的时间空闲下来喝咖啡）。对于电路交换，在所有的时间内必须为每个用户预留 100kbps。例如，对于电路交换的 TDM，如果一个 1s 的帧被划分为 10 个时隙，每个时隙为 100ms，则每帧将为每个用户分配一个时隙。

因此，该电路交换链路仅能支持 10 (=1Mbps/100kbps) 个并发的用户。对于分组交换，一个特定用户活跃的概率是 0.1 (即 10%)。如果有 35 个用户，有 11 或更多个并发活跃用户的概率大约是 0.0004。(课后习题 P8 概述如何得到这个概率值。) 当有 10 个或更少并发用户 (以概率 0.9996 发生) 时，到达的聚合数据速率小于或等于该链路的输出速率 1Mbps。因此，当有 10 个或更少的活跃用户时，通过该链路的分组流基本上没有时延，这与电路交换的情况一样。当同时活跃用户超过 10 个时，分组的聚合到达速率超过该链路的输出容量，则输出队列将开始变长。(一直增长到聚合输入速率重新低于 1Mbps，此后该队列长度才会减少。) 因为在本例子中同时活跃用户超过 10 个的概率极小，分组交换差不多总是提供了与电路交换相同的性能，并且允许在用户数量是其 3 倍时情况也是如此。

我们现在考虑第二个简单的例子。假定有 10 个用户，某个用户突然产生 1000 个 1000 比特的分组，而其他用户则保持静默，不产生分组。在每帧具有 10 个时隙并且每个时隙包含 1000 比特的 TDM 电路交换情况下，活跃用户仅能使用每帧中的一个时隙来传输数据，而每个帧中剩余的 9 个时隙保持空闲。该活跃用户传输完所有 10^6 比特数据需要 10s 的时间。在分组交换情况下，活跃用户能够连续地以 1Mbps 的全部链路速率发送其分组，因为没有其他用户产生分组与该活跃用户的分组进行复用。在此情况下，该活跃用户的所

有数据将在1s内发送完毕。

上面的例子从两个方面表明了分组交换的性能能够优于电路交换的性能。这些例子也强调了在多个数据流之间共享链路传输速率的两种形式的关键差异。电路交换不考虑需求，而预先分配了传输链路的使用，这使得已分配而并不需要的链路时间未被利用。另一方面，分组交换按需分配链路使用。链路传输能力将在所有需要在链路上传输分组的用户之间逐分组地被共享。

虽然分组交换和电路交换在今天的电信网络中都是普遍采用的方式，但趋势无疑是朝着分组交换方向发展。甚至许多今天的电路交换电话网正在缓慢地向分组交换迁移。特别是，电话网经常在昂贵的海外电话部分使用分组交换。

1.3.3　网络的网络

我们在前面看到，端系统（PC、智能手机、Web服务器、电子邮件服务器等）经过一个接入ISP与因特网相连。该接入ISP能够提供有线或无线连接，使用了包括DSL、电缆、FTTH、WiFi和蜂窝等多种接入技术。值得注意的是，接入ISP不必是电信局或电缆公司，相反，它能够是如大学（为学生、教职员工和从业人员提供因特网接入）或公司（为其雇员提供接入）这样的单位。但让端用户和内容提供商连接到接入ISP仅解决了连接难题中的很小一部分，因为因特网是由数以亿计的用户构成的。要解决这个难题，接入ISP自身必须互联。通过创建网络的网络可以做到这一点，理解这个短语是理解因特网的关键。

年复一年，构成因特网的“网络的网络”已经演化成为一个非常复杂的结构。这种演化很大部分是由经济和国家策略驱动的，而不是由性能考虑驱动的。为了理解今天的因特网的网络结构，我们以逐步递进方式建造一系列网络结构，其中的每个新结构都更好地接近现在的复杂因特网。回顾前面互联接入ISP的中心目标，是使所有端系统能够彼此发送分组。一种幼稚的方法是使每个接入ISP直接与每个其他接入ISP连接。当然，这样的网状设计对于接入ISP费用太高，因为这将要求每个接入ISP要与世界上数十万个其他接入ISP有一条单独的通信链路。

我们的第一个网络结构即网络结构1，用单一的全球传输ISP互联所有接入ISP。我们假想的全球传输ISP是一个由路由器和通信链路构成的网络，该网络不仅跨越全球，而且至少具有一台路由器靠近数十万接入ISP中的每一个。当然，对于全球传输ISP，建造这样一个大规模的网络将耗资巨大。为了有利可图，自然要向每个连接的接入ISP收费，其价格反映（并不一定正比于）一个接入ISP经过全球ISP交换的流量大小。因为接入ISP向全球传输ISP付费，故接入ISP被认为是**客户**（customer），而全球传输ISP被认为是**提供商**（provider）。

如果某个公司建立并运营一个可赢利的全球传输ISP，其他公司建立自己的全球传输ISP并与最初的全球传输ISP竞争则是一件自然的事。这导致了网络结构2，它由数十万接入ISP和多个全球传输ISP组成。接入ISP无疑喜欢网络结构2胜过喜欢网络结构1，因为它们现在能够根据价格和服务因素在多个竞争的全球传输提供商之间进行选择。然而，值得注意的是，这些全球传输ISP之间必须是互联的；不然的话，与某个全球传输ISP连接的接入ISP将不能与连接到其他全球传输ISP的接入ISP进行通信。

刚才描述的网络结构2是一种两层的等级结构，其中全球传输提供商位于顶层，而接入ISP位于底层。这假设了全球传输ISP不仅能够接近每个接入ISP，而且发现经济上也

希望这样做。现实中，尽管某些 ISP 确实具有令人印象深刻的全球覆盖，并且确实直接与许多接入 ISP 连接，但世界上没有哪个 ISP 是无处不在的。相反，在任何给定的区域，可能有一个**区域 ISP**（regional ISP），区域中的接入 ISP 与之连接。每个区域 ISP 则与**第一层 ISP**（tier-1 ISP）连接。第一层 ISP 类似于我们假想的全球传输 ISP，尽管它不是在世界上每个城市中都存在，但它确实存在。有大约十几个第一层 ISP，包括 Level 3 Communications、AT&T、Sprint 和 NTT。有趣的是，没有组织正式认可第一层状态。俗话说：如果必须问你是否是一个组织的成员，你可能不是。

再来讨论这个网络的网络，不仅有多个竞争的第一层 ISP，而且在一个区域可能有多个竞争的区域 ISP。在这样的等级结构中，每个接入 ISP 向其连接的区域 ISP 支付费用，并且每个区域 ISP 向它连接的第一层 ISP 支付费用。（一个接入 ISP 也能直接与第一层 ISP 连接，这样它就向第一层 ISP 付费。）因此，在这个等级结构的每一层，都有客户 - 提供商关系。值得注意的是，第一层 ISP 不向任何人付费，因为它们位于该等级结构的顶部。更为复杂的情况是，在某些区域，可能有较大的区域 ISP（可能跨越整个国家），该区域中较小的区域 ISP 与之相连，较大的区域 ISP 则与第一层 ISP 连接。例如，在中国，每个城市有接入 ISP，它们与省级 ISP 连接，省级 ISP 又与国家级 ISP 连接，国家级 ISP 最终与第一层 ISP 连接［Tian 2012］。这个多层等级结构仍然仅仅是今天因特网的粗略近似，我们称它为*网络结构 3*。

为了建造一个与今天的因特网更为相似的网络，我们必须在等级化网络结构 3 上增加存在点（Point of Presence，PoP）、多宿、对等和因特网交换点。PoP 存在于等级结构的所有层次，但底层（接入 ISP）等级除外。一个 PoP 只是提供商网络中的一台或多台路由器（在相同位置）群组，其中客户 ISP 能够与提供商 ISP 连接。对于要与提供商 PoP 连接的客户网络，它能从第三方电信提供商租用高速链路将它的路由器之一直接连接到位于该 PoP 的一台路由器。任何 ISP（除了第一层 ISP）可以选择**多宿**（multi-home），即可以与两个或更多提供商 ISP 连接。例如，一个接入 ISP 可能与两个区域 ISP 多宿，既可以与两个区域 ISP 多宿，也可以与一个第一层 ISP 多宿。当一个 ISP 多宿时，即使它的提供商之一出现故障，它仍然能够继续发送和接收分组。

正如我们刚才学习的，客户 ISP 向它们的提供商 ISP 付费以获得全球因特网互联能力。客户 ISP 支付给提供商 ISP 的费用数额反映了它通过提供商交换的通信流量。为了减少这些费用，位于相同等级结构层次的邻近一对 ISP 能够**对等**（peer），也就是说，能够直接将它们的网络连到一起，使它们之间的所有流量经直接连接而不是通过上游的中间 ISP 传输。当两个 ISP 对等时，通常不进行结算，即任一个 ISP 不向其对等付费。如前面提到的那样，第一层 ISP 也与另一个第一层 ISP 对等，它们之间无结算。对于对等和客户 - 提供商关系的讨论，［Van der Berg 2008］是一本不错的读物。沿着这些相同路线，第三方公司能够创建一个**因特网交换点**（Internet Exchange Point，IXP），IXP 是一个汇合点，多个 ISP 能够在这里一起对等。IXP 通常位于一个有自己的交换机的独立建筑物中［Ager 2012］，在今天的因特网中有 400 多个 IXP［IXP List 2016］。我们称这个生态系统为*网络结构 4*——由接入 ISP、区域 ISP、第一层 ISP、PoP、多宿、对等和 IXP 组成。

我们现在最终到达了网络结构 5，它描述了现今的因特网。在图 1-15 中显示了网络结构 5，它通过在网络结构 4 顶部增加**内容提供商网络**（content provider network）构建而成。谷歌是当前这样的内容提供商网络的一个突出例子。在本书写作之时，谷歌估计有50 ~ 100 个数据中心分布于北美、欧洲、亚洲、南美和澳大利亚。其中的某些数据中心容纳了

超过十万台的服务器，而另一些数据中心则较小，仅容纳数百台服务器。谷歌数据中心都经过专用的 TCP/IP 网络互联，该网络跨越全球，不过独立于公共因特网。重要的是，谷歌专用网络仅承载出入谷歌服务器的流量。如图 1-15 所示，谷歌专用网络通过与较低层 ISP 对等（无结算），尝试“绕过”因特网的较高层，采用的方式可以是直接与它们连接，或者在 IXP 处与它们连接［Labovitz 2010］。然而，因为许多接入 ISP 仍然仅能通过第一层网络的传输到达，所以谷歌网络也与第一层 ISP 连接，并就与这些 ISP 交换的流量向它们付费。通过创建自己的网络，内容提供商不仅减少了向顶层 ISP 支付的费用，而且对其服务最终如何交付给端用户有了更多的控制。谷歌的网络基础设施在 2.6 节中进行了详细描述。

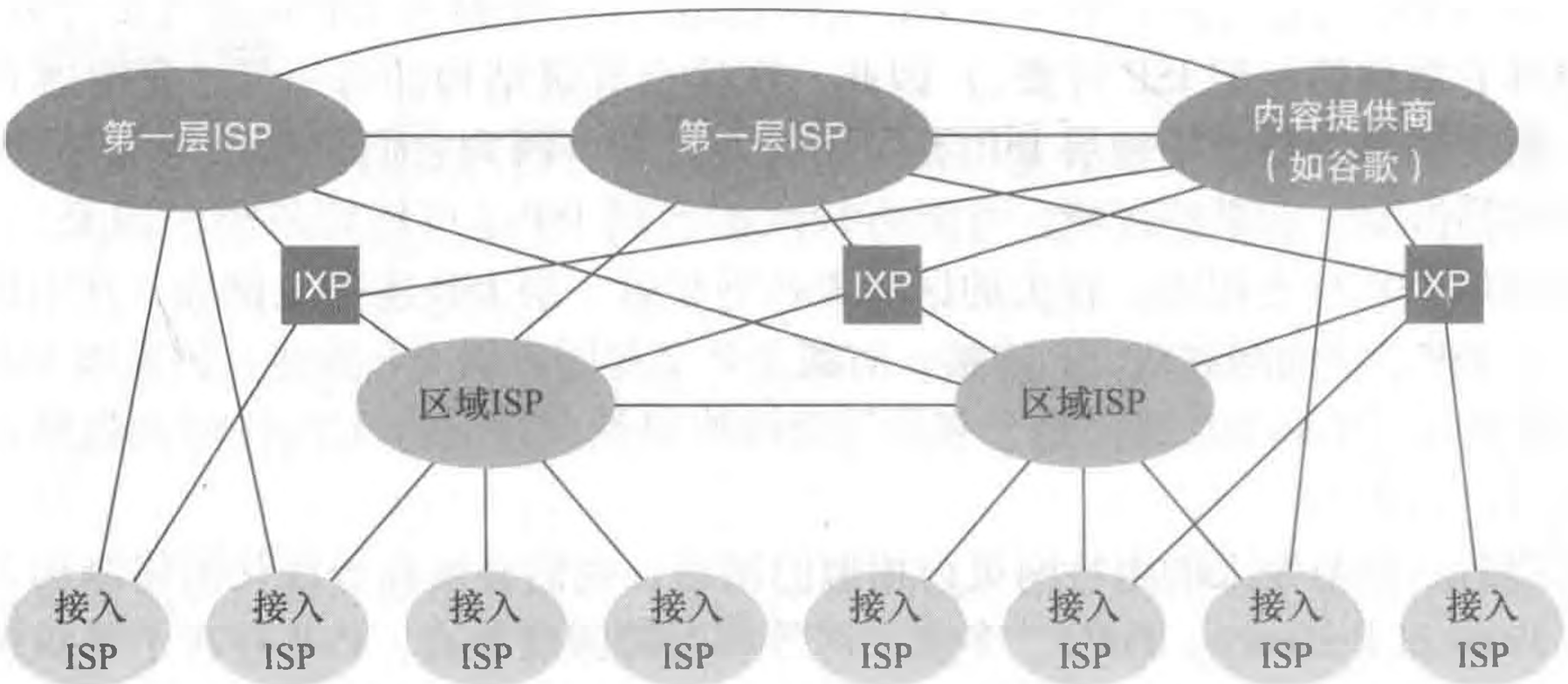

图 1-15　ISP 的互联

总结一下，今天的因特网是一个网络的网络，其结构复杂，由十多个第一层 ISP 和数十万个较低层 ISP 组成。ISP 覆盖的区域多种多样，有些跨越多个大洲和大洋，有些限于狭窄的地理区域。较低层的 ISP 与较高层的 ISP 相连，较高层 ISP 彼此互联。用户和内容提供商是较低层 ISP 的客户，较低层 ISP 是较高层 ISP 的客户。近年来，主要的内容提供商也已经创建自己的网络，直接在可能的地方与较低层 ISP 互联。

1.4　分组交换网中的时延、丢包和吞吐量

回想在 1.1 节中我们讲过，因特网能够看成是一种基础设施，该基础设施为运行在端系统上的分布式应用提供服务。在理想情况下，我们希望因特网服务能够在任意两个端系统之间随心所欲地瞬间移动数据而没有任何数据丢失。然而，这是一个极高的目标，实践中难以达到。与之相反，计算机网络必定要限制在端系统之间的吞吐量（每秒能够传送的数据量），在端系统之间引入时延，而且实际上也会丢失分组。一方面，现实世界的物理定律引入的时延、丢包并限制吞吐量是不幸的。而另一方面，因为计算机网络存在这些问题，围绕如何去处理这些问题有许多令人着迷的话题，多得足以开设一门有关计算机网络方面的课程，可以做上千篇博士论文！在本节中，我们将开始研究和量化计算机网络中的时延、丢包和吞吐量等问题。

1.4.1　分组交换网中的时延概述

前面讲过，分组从一台主机（源）出发，通过一系列路由器传输，在另一台主机

（目的地）中结束它的历程。当分组从一个节点（主机或路由器）沿着这条路径到后继节点（主机或路由器），该分组在沿途的每个节点经受了几种不同类型的时延。这些时延最为重要的是**节点处理时延**（nodal processing delay）、**排队时延**（queuing delay）、**传输时延**（transmission delay）和**传播时延**（propagation delay），这些时延总体累加起来是**节点总时延**（total nodal delay）。许多因特网应用，如搜索、Web 浏览、电子邮件、地图、即时讯息和 IP 语音，它们的性能受网络时延的影响很大。为了深入理解分组交换和计算机网络，我们必须理解这些时延的性质和重要性。

时延的类型

我们来探讨一下图 1-16 环境中的这些时延。作为源和目的地之间的端到端路由的一部分，一个分组从上游节点通过路由器 A 向路由器 B 发送。我们的目标是在路由器 A 刻画出节点时延。值得注意的是，路由器 A 具有通往路由器 B 的出链路。该链路前面有一个队列（也称为缓存）。当分组从上游节点到达路由器 A 时，路由器 A 检查该分组的首部以决定它的适当出链路，并将该分组导向该链路。在这个例子中，对该分组的出链路是通向路由器 B 的那条链路。仅当在该链路没有其他分组正在传输并且没有其他分组排在该队列前面时，才能在这条链路上传输该分组；如果该链路当前正繁忙或有其他分组已经在该链路上排队，则新到达的分组将加入排队。

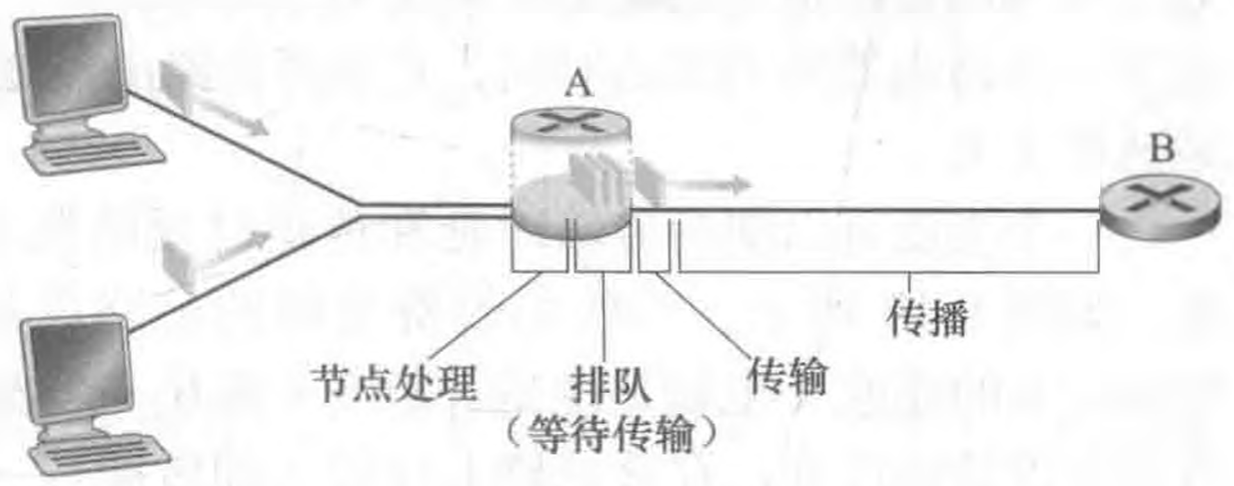

图 1-16　路由器 A 的节点时延

（1）处理时延

检查分组首部和决定将该分组导向何处所需要的时间是**处理时延**的一部分。处理时延也能够包括其他因素，如检查比特级别的差错所需要的时间，该差错出现在从上游节点向路由器 A 传输这些分组比特的过程中。高速路由器的处理时延通常是微秒或更低的数量级。在这种节点处理之后，路由器将该分组引向通往路由器 B 链路之前的队列。（在第 4 章中，我们将研究路由器运行的细节。）

（2）排队时延

在队列中，当分组在链路上等待传输时，它经受**排队时延**。一个特定分组的排队时延长度将取决于先期到达的正在排队等待向链路传输的分组数量。如果该队列是空的，并且当前没有其他分组正在传输，则该分组的排队时延为 0。另一方面，如果流量很大，并且许多其他分组也在等待传输，该排队时延将很长。我们将很快看到，到达分组期待发现的分组数量是到达该队列的流量的强度和性质的函数。实际的排队时延可以是毫秒到微秒量级。

（3）传输时延

假定分组以先到先服务方式传输——这在分组交换网中是常见的方式，仅当所有已经到达的分组被传输后，才能传输刚到达的分组。用 L 比特表示该分组的长度，用 R bps（即 b/s）表示从路由器 A 到路由器 B 的链路传输速率。例如，对于一条 10Mbps 的以太网链路，速率 $R = 10$Mbps；对于 100Mbps 的以太网链路，速率 $R = 100$Mbps。**传输时延**是 L/R。这是将所有分组的比特推向链路（即传输，或者说发射）所需要的时间。实际的传输时延通常在毫秒到微秒量级。

（4）传播时延

一旦一个比特被推向链路，该比特需要向路由器 B 传播。从该链路的起点到路由器 B 传播所需要的时间是**传播时延**。该比特以该链路的传播速率传播。该传播速率取决于该链路的物理媒体（即光纤、双绞铜线等），其速率范围是 $2\times10^8\sim3\times10^8$ m/s，这等于或略小于光速。该传播时延等于两台路由器之间的距离除以传播速率。即传播时延是 d/s，其中 d 是路由器 A 和路由器 B 之间的距离，s 是该链路的传播速率。一旦该分组的最后一个比特传播到节点 B，该比特及前面的所有比特被存储于路由器 B。整个过程将随着路由器 B 执行转发而持续下去。在广域网中，传播时延为毫秒量级。

（5）传输时延和传播时延的比较

计算机网络领域的新手有时难以理解传输时延和传播时延之间的差异。该差异是微妙而重要的。传输时延是路由器推出分组所需要的时间，它是分组长度和链路传输速率的函数，而与两台路由器之间的距离无关。另一方面，传播时延是一个比特从一台路由器传播到另一台路由器所需要的时间，它是两台路由器之间距离的函数，而与分组长度或链路传输速率无关。

一个类比可以阐明传输时延和传播时延的概念。考虑一条公路每 100km 有一个收费站，如图 1-17 所示。可认为收费站间的公路段是链路，收费站是路由器。假定汽车以 100km/h 的速度（也就是说当一辆汽车离开一个收费站时，它立即加速到 100km/h 并在收费站间维持该速度）在该公路上行驶（即传播）。假定这时有 10 辆汽车作为一个车队在行驶，并且这 10 辆汽车以固定的顺序互相跟随。可以认为每辆汽车是一个比特，该车队是一个分组。同时假定每个收费站以每辆车 12s 的速度服务（即传输）一辆汽车，并且由于时间是深夜，因此该车队是公路上唯一一批汽车。最后，假定无论该车队的第一辆汽车何时到达收费站，它在入口处等待，直到其他 9 辆汽车到达并整队依次前行。（因此，整个车队在它能够“转发”之前，必须存储在收费站。）收费站将整个车队推向公路所需要的时间是（10 辆车）/（5 辆车/min）= 2min。该时间类比于一台路由器中的传输时延。一辆汽车从一个收费站出口行驶到下一个收费站所需要的时间是 100km/（100km/h）= 1h。这个时间类比于传播时延。因此，从该车队存储在收费站前到该车队存储在下一个收费站前的时间是“传输时延”与“传播时间”总和，在本例中为 62min。

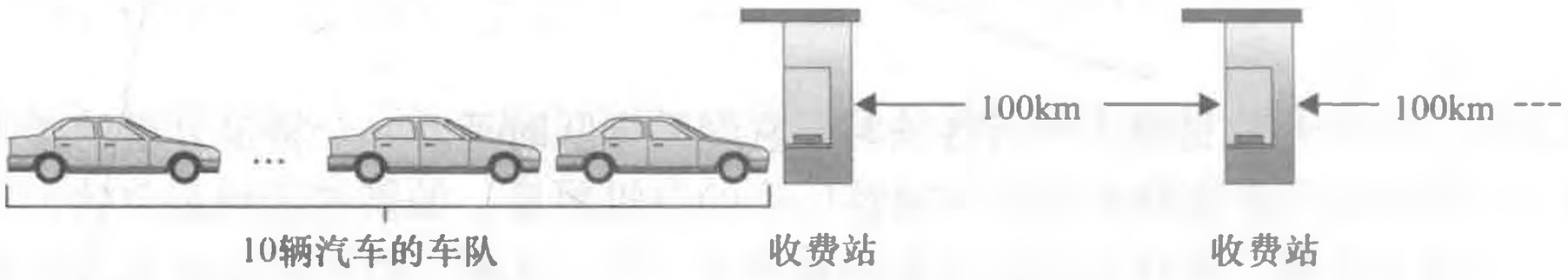

图 1-17　车队的类比

我们更深入地探讨一下这个类比。如果收费站对车队的服务时间大于汽车在收费站之间行驶的时间，将会发生什么情况呢？例如，假定现在汽车是以 1000km/h 的速率行驶，收费站是以每分钟一辆汽车的速率为汽车服务。则汽车在两个收费站之间的行驶时延是 6min，收费站为车队服务的时间是 10min。在此情况下，在该车队中的最后几辆汽车离开第一个收费站之前，该车队中前面的几辆汽车将会达到第二个收费站。这种情况在分组交换网中也会发生，一个分组中的前几个比特到达了一台路由器，而该分组中许多余下的比特仍然在前面的路由器中等待传输。

如果说一图胜千言的话，则一个动画必定胜百万言。与本书配套的 Web 网站提供了

一个交互式 Java 小程序，它很好地展现及对比了传输时延和传播时延。我们极力推荐读者访问该 Java 小程序。[Smith 2009] 也提供了可读性很好的有关传播、排队和传输时延的讨论。

如果令 d_{proc}、d_{queue}、d_{trans}和 d_{prop}分别表示处理时延、排队时延、传输时延和传播时延，则节点的总时延由下式给定：

$$d_{nodal} = d_{proc} + d_{queue} + d_{trans} + d_{prop}$$

这些时延成分所起的作用可能会有很大的不同。例如，对于连接两台位于同一个大学校园的路由器的链路而言，d_{prop}可能是微不足道的（例如，几微秒）；然而，对于由同步卫星链路互联的两台路由器来说，d_{prop}是几百毫秒，能够成为 d_{nodal}中的主要成分。类似地，d_{trans}的影响可能是微不足道的，也可能是很大的。通常对于 10Mbps 和更高的传输速率（例如，对于 LAN）的信道而言，它的影响是微不足道的；然而，对于通过低速拨号调制解调器链路发送的长因特网分组而言，可能是数百毫秒。处理时延 d_{proc}通常是微不足道的；然而，它对一台路由器的最大吞吐量有重要影响，最大吞吐量是一台路由器能够转发分组的最大速率。

1.4.2　排队时延和丢包

节点时延的最为复杂和有趣的成分是排队时延 d_{queue}。事实上，排队时延在计算机网络中的重要程度及人们对它感兴趣的程度，从发表的数以千计的论文和大量专著的情况可见一斑 [Bertsekas 1991；Daigle 1991；Kleinrock 1975，1976；Ross 1995]。我们这里仅给出有关排队时延的总体的、直觉的讨论；求知欲强的读者可能要浏览某些书籍（或者最终写有关这方面的博士论文）。与其他 3 项时延（即 d_{proc}、d_{trans}和 d_{prop}）不同的是，排队时延对不同的分组可能是不同的。例如，如果 10 个分组同时到达空队列，传输的第一个分组没有排队时延，而传输的最后一个分组将经受相对大的排队时延（这时它要等待其他 9 个分组被传输）。因此，当表征排队时延时，人们通常使用统计量来度量，如平均排队时延、排队时延的方差和排队时延超过某些特定值的概率。

什么时候排队时延大，什么时候又不大呢？该问题的答案很大程度取决于流量到达该队列的速率、链路的传输速率和到达流量的性质，即流量是周期性到达还是以突发形式到达。为了更深入地领会某些要点，令 a 表示分组到达队列的平均速率（a 的单位是分组/秒，即 pkt/s）。前面讲过 R 是传输速率，即从队列中推出比特的速率（以 bps 即 b/s 为单位）。为了简单起见，也假定所有分组都是由 L 比特组成的。则比特到达队列的平均速率是 La bps。最后，假定该队列非常大，因此它基本能容纳无限数量的比特。比率 La/R 被称为**流量强度**（traffic intensity），它在估计排队时延的范围方面经常起着重要的作用。如果 $La/R > 1$，则比特到达队列的平均速率超过从该队列传输出去的速率。在这种不幸的情况下，该队列趋向于无限增加，并且排队时延将趋向无穷大！因此，流量工程中的一条金科玉律是：*设计系统时流量强度不能大于 1*。

现在考虑 $La/R \leqslant 1$ 时的情况。这时，到达流量的性质影响排队时延。例如，如果分组周期性到达，即每 L/R 秒到达一个分组，则每个分组将到达一个空队列中，不会有排队时延。另一方面，如果分组以突发形式到达而不是周期性到达，则可能会有很大的平均排队时延。例如，假定每 $(L/R)N$ 秒同时到达 N 个分组。则传输的第一个分组没有排队时延；传输的第二个分组就有 L/R 秒的排队时延；更为一般地，第 n 个传输的分组具有 $(n-1)$