议，最后传递给链路层协议。所有这些都发生在客户端设备上。此时，该请求被传播到本地网络（图中标记为网络1）上。这个请求通过互联网，从一个网络传到另一个网络。本例中，路由器A把该请求从网络1路由到网络2，路由器B把该请求从网络2路由到网络3。当请求到达目标服务器后，它将沿着网络协议向上传递，从链路层协议开始，一直到应用层协议结束。该服务器上的一个进程接收该请求，根据客户端最初使用的应用层协议进行格式化。服务器进程解释该请求，并以适当的方式进行响应。

现在我们从底层开始，看看其中的每一层。

11.2.1 链路层

互联网协议套件的底层就是链路层。主机之间的物理和逻辑连接被称为网络链路。同一个网络上的设备使用链路层协议相互通信。链路上的每个设备都有一个网络地址能唯一地标识它。对于许多链路层协议而言，这个地址被称为媒体访问控制（Media Access Control，MAC）地址。链路层数据被划分成小单元，称为“帧”，每帧都包含一个描述该帧的帧标题、数据的有效负载，以及最后用于检测错误的帧尾，如图11-6所示。

图11-6 链路层帧

帧标题包含源MAC地址和目的MAC地址。帧标题还包含对帧数据部分所携带数据的类型描述。

如果你家里有Wi-Fi网络，那么Wi-Fi就是你网络上主机之间的链路。由IEEE 802.11规范定义的Wi-Fi协议不知道、也不关心在无线网络上发送的数据是什么类型；它只启用允许设备之间的通信。每个连接到Wi-Fi网络的设备都有一个MAC地址，并接收发送到其地址的帧。MAC地址仅在本地网络上可用，远程网络上的计算机不能直接向本地网络上的MAC地址发送数据。

另一个值得注意的链路层技术是以太网（Ethernet），用于有线物理连接。以太网由IEEE 802.3标准定义。以太网一般使用内部有一对铜线的电缆，其末端是通常被称为RJ45的接头，如图11-7所示。

连接到互联网的所有设备都参与到链路层中。这是必需的，因为链路层提供的是到本地网络的连接（不论是有线还是无线）。主机（比如笔记本计算机或智能手机）参与所有的层次，但是某些网络设备只在链路层操作。最基本的例子就是集线器（hub）。网络集线器是一种网络设备，它连接本地网络上的多个设备，无须具有对正在发送的帧的智能。简单的集线器可以对连接的设备提供多个以太网端口。它只是把在一个物理端口接收的每个帧传送到其他所有的端口。更智能一些的链路层设备是网络交换机，它要检查所接收帧中的MAC地址，并把这些帧发送到具有目标MAC地址的设备所连接的物理端口。

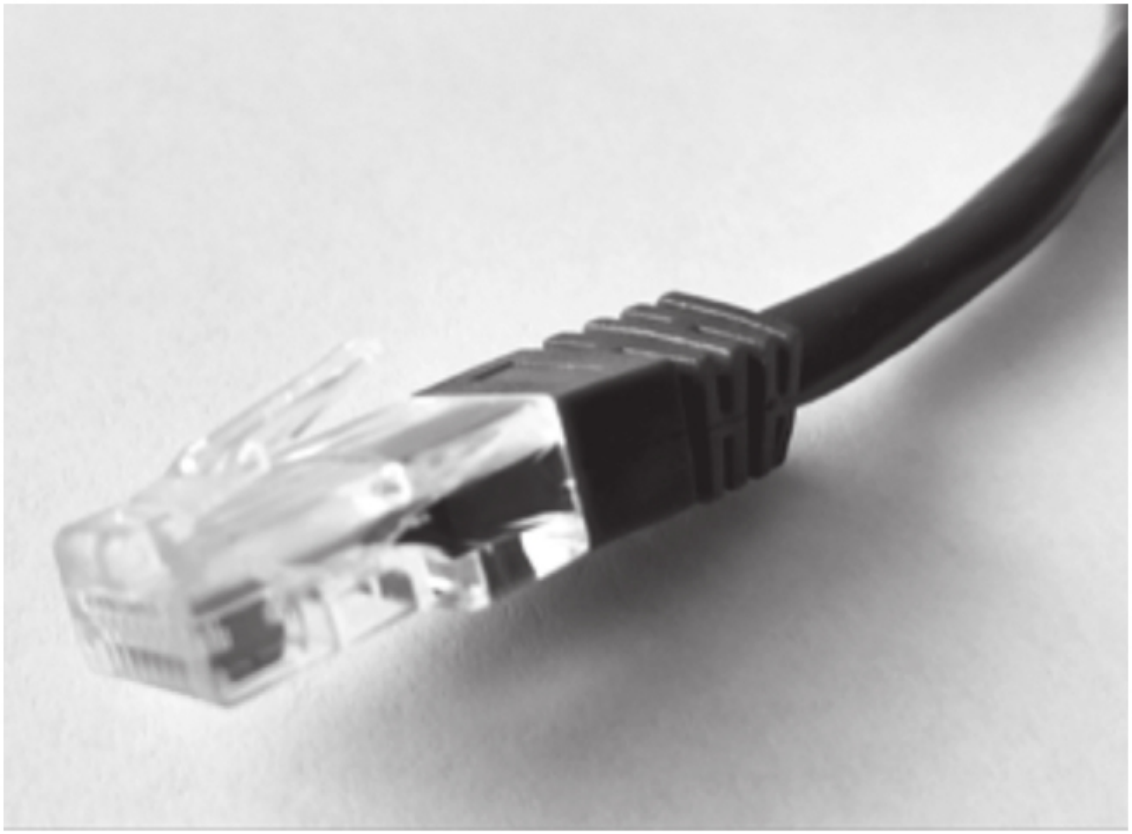

图11-7 常用于以太网的电缆

注意

请参阅设计29查看链路层设备和MAC地址。

11.2.2 网络层

网络层使得数据在本地网络之外传送。本层使用的主要协议简称为互联网协议（IP）。它支持路由，即为网络间传输的数据决定路径的过程。互联网上的每个主机都被分配了一个IP地址，它是一个唯一标识全球互联网上主机的数字。主机也可能拥有不直接在互联网上公开的私有IP地址。IP地址一般由本地网络上的服务器分配，当设备连接到新网络时，其IP地址通常会改变。稍后，我们将更详细地介绍地址分配和私有IP地址。

在网络层上传送的数据被称为包，它被封装在链路层帧中。图11-8展示了"包在帧数据部分中"的含义。

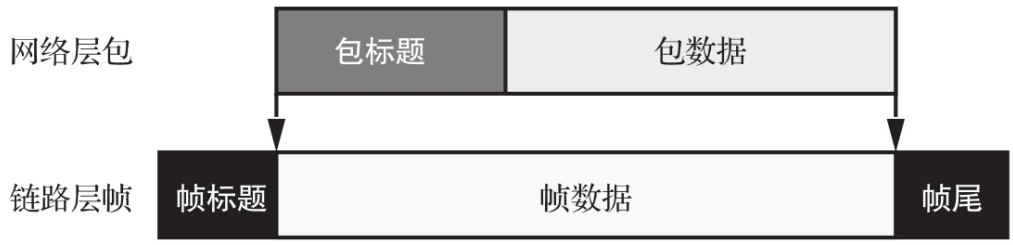

图11-8 帧数据部分中的包

IP包标题含有源IP地址和目标IP地址。包标题还包含了包的描述信息，比如使用的IP版本以及包标题长度。IP包的数据部分包含IP层携带的有效负载。

如今互联网使用了两个版本的互联网协议。互联网协议版本4（IPv4）是主要使用的版本，另一活跃版本是互联网协议版本6（IPv6）。你可能会好奇：没有IPv5吗？这样的协议从未存在过，但是一个被称为互联网流协议的实验协议把它的IP版本称为IPv5，所以在开发IPv4的继任者时跳过了IPv5。IPv4与IPv6的一个显著区别是IP地址的大小。IPv4地址长32位，而

IPv6地址长128位。这个差异使得IPv6的地址数量要多得多。地址大小的变化有助于解决IPv4地址相对短缺的问题。本书中，我们关注的是IPv4地址（简称IP地址），因为它们仍然是如今互联网的主要寻址方式。

32位IP地址通常用点分十进制表示法来显示，这意味着32位被分成四个8位组，每组中的8位都用十进制（而不是十六进制或二进制）显示。四个十进制数用句点（.）分隔。例如，一个IP地址用点分十进制表示为192.168.1.23。每个8位十进制数可以称为一个8位字节（octet）。

对于连接到同一个本地网络的计算机，其IP地址具有相同的前导位，称为位于同一子网（subnet）。位于同一子网的计算机，相互之间能在链路层上直接通信，因为它们在同一物理网络上运行。位于不同子网的计算机必须通过路由器发送其流量，路由器是连接子网的设备，它在网络层上运行。

子网把IP地址划分为两个部分：网络前缀和主机标识符。前者由同一子网上的所有设备共享，后者对于子网上的主机来说是唯一的。网络前缀所包含的位数随网络配置变化。

让我们看个例子。假设一个子网使用了24位的网络前缀，留下8位表示主机。假设该子网上的主机使用之前的示例IP地址192.168.1.23。给定这个IP地址和网络前缀，该IP地址的划分如图11-9所示。

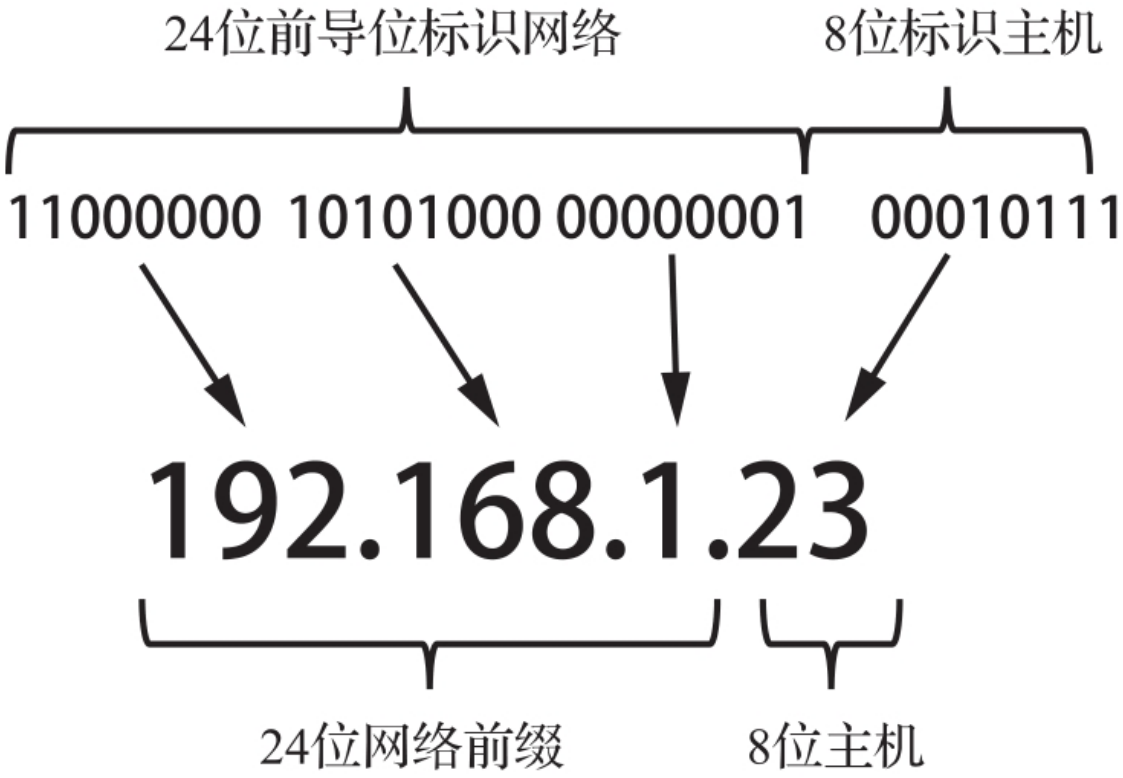

图11-9 使用24位网络前缀的IP地址示例

在这个例子中，本地子网上所有主机的IP地址都以192.168.1开头。每个主机的最后一个8位字节具有不同的值，其中的23分配给此例中的特定主机。这个例子使用24位前缀长度，表示前缀与IP地址的前3个8位字节对齐。这是个很好的例子，但前缀长度并不总是与8位字节的边界对齐。例如，25位前缀会包含最后一个8位字节的第一位，只留下7位标识主机。

为网络前缀保留的位数通常有两种表示方式。无类别域间路由（Classless Inter-Domain Routing，CIDR）表示法列出一个IP地址，其后跟一个斜杠（/），然后是网络前缀使用的位数。在我们的例子中，应表示为192.168.1.23/24。另一种表示前缀位数的常用方法是采用子网掩码——一个32位的数字，网络前缀部分中的每个位用二进制1表示，主机号部分的每个位用0表示。子网掩码也写作点分十进制的形式，所以我们的24位网络前缀例子的子网掩码为255.255.255.0，如图11-10所示。

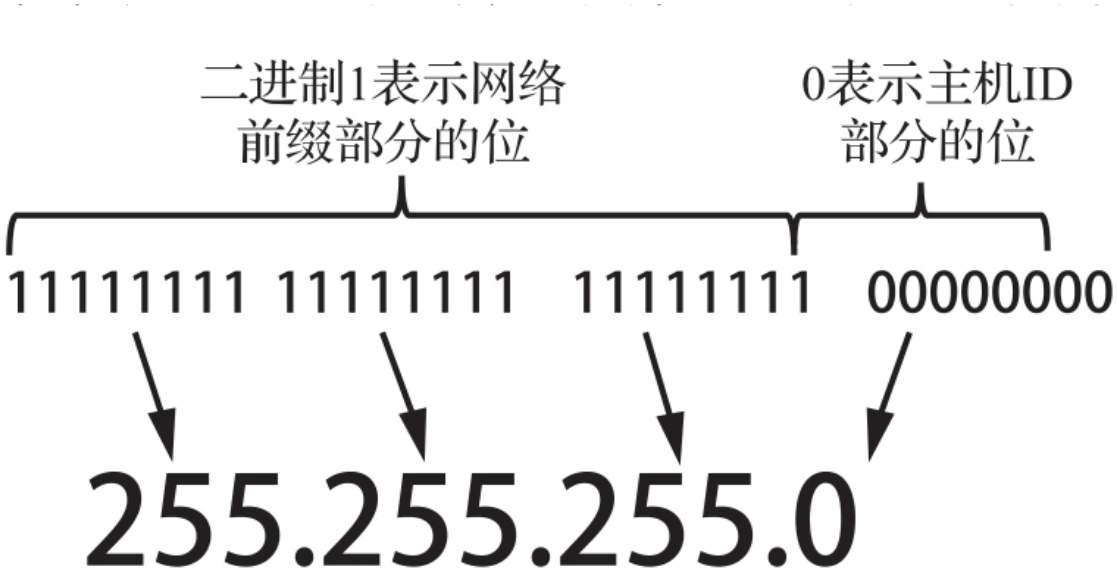

图11-10 表示为子网掩码的24位网络前缀

让我们看看这在实践中为什么有用。假设计算机的IP地址是192.168.0.133，子网掩码是255.255.255.244，用CIDR表示法表示为192.168.0.133/27。假设此计算机想连接到另一台IP地址为192.168.0.84的计算机。如前所述，如果两台计算机在同一个子网上，则它们可以直接通信；如果不在同一个子网上，它们必须通过路由器来通信。因此，该计算机必须确定另一台计算机是否在同一个子网上。如何做到这一点呢?

对IP地址和它的子网掩码执行按位逻辑AND运算，就可以生成子网中的第一个地址。第一个地址（其中的主机位全为0）用作子网自身的标识符。这通常被称为网络ID。共享网络ID的两个计算机在同一个子网上。主机可以对自己的IP地址和它想要连接的IP地址执行这个AND操作，以查看它们是否共享同一个网络ID，从而判断是否在同一个子网上。让我们用示例计算机的IP地址来试一下，如下所示：

```
  IP = 192.168.0.133   = 11000000.10101000.00000000.10000101
MASK = 255.255.255.224 = 11111111.11111111.11111111.11100000
 AND = 192.168.0.128   = 11000000.10101000.00000000.10000000 = The network ID
```

现在，对我们示例中的第二台计算机执行同样的操作：

```
  IP = 192.168.0.84    = 11000000.10101000.00000000.01010100
MASK = 255.255.255.224 = 11111111.11111111.11111111.11100000
 AND = 192.168.0.64    = 11000000.10101000.00000000.01000000 = The network ID
```

从本例中可以看到，这个操作产生了两个不同的网络ID（192.168.0.128和192.168.0.64）。这表示第二台计算机与第一台计算机不在同一个子网上。若要通信，这两台计算机需要通过路由器来发送它们的信息，这个路由器连接两个子网。

练习11-1：哪些I P地址在同一个子网上?

IP地址192.168.0.200和示例计算机在同一个子网上吗？假设示例计算机的IP地址是192.168.0.133，且子网掩码为255.255.255.224。

还有另一种查看方式：网络前缀描述了子网上能使用的地址范围。该范围中的第一个地址被定义为网络前缀加上全二进制0的主机标识符。继续使用IP地址为192.168.0.133的示例计算机，其子网上的第一个地址是192.168.0.128。该范围中的最后一个地址是网络前缀加上全二进制1的主机标识符。在我们的例子中，这个地址是192.168.0.159。第一个地址和最后一个地址有特殊含义——第一个标识网络，最后一个表示广播地址（用于向子网上的全部主机发送消息）。两者之间的所有地址都可以用于子网上的主机。我们的示例IP地址192.168.0.133显然在这个范围（从192.168.0.128到192.168.0.159）之内，而另一个IP地址为192.168.0.84的计算机则不在这个范围中。

你还可以用为主机标识符保留的位数来确定子网上有多少IP地址可供主机使用。在我们的例子中，27位用于网络前缀，剩下5位用于主机标识符。这5位提供给我们32个可用主机地址，因为2^5等于32。但是，正如前面所说的，第一个和最后一个地址有特殊用途，所以使用这个网络前缀的话，实际上只有30台主机能被标识。这与我们前面的发现是一致的：第一个主机标识符是128，128+32等于160，它是下一个子网的第一个地址，所以159是该地址范围内的最后一个主机。

注意

请参阅设计30用自己的Raspberry Pi查看网络层。

11.2.3 传输层

传输层为应用程序收发数据提供可用的通信通道。有两种常用的传输层协议：传输控制协议（TCP）和用户数据报协议（User Datagram Protocol，UDP）。TCP提供两个主机之间的可靠连接。它确保错误数最小，数据按序到达，重发丢失数据等。用TCP发送的数据被称为段（segment）。UDP是一种"尽力而为"的协议，意思是说它的交付是不可靠的。当速度比可靠性更重要时，UDP是首选。用UDP发送的数据被称为数据报。这两个协议都有自己的适用条件，但为了简单起见，本章剩余部分只介绍TCP。图11-11展示了TCP段在包的数据部分中的含义，而包又在帧的数据部分。

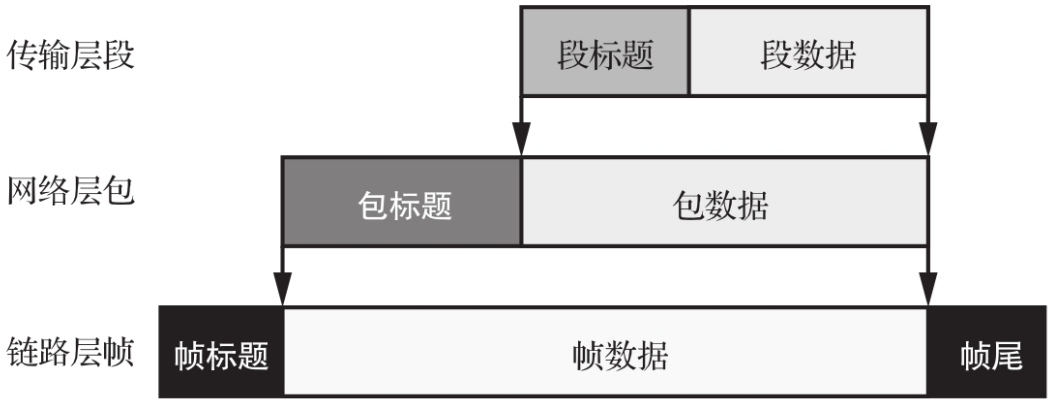

图11-11 TCP段包含在IP包的数据部分中

如前所述，链路层在帧标题中包含了目标MAC地址，以识别本地网络接口，网络层在包标题中包含了目标IP地址，以识别互联网上的主机。这些信息足以让一个数据包到达互联网上的特定设备。数据包到达其目标主机后，传输层段标题中包含了目标网络端口号，该端口号标识了将要接收数据的特定服务或进程。单个IP地址的主机可以有多个活跃端口，每个端口都用于执行不同类型的网络活动。

打个比方，IP地址就像是办公楼的街道地址，网络端口号就像是该办公楼中工作人员的办公室号码。IP地址唯一地标识主机，就像街道地址唯一地标识办公楼一样。使用互联网协议，数据包可以被传递给某个主机，就像包裹可以被送达办公楼一样。但是，数据包到达计算机后，操作系统必须决定如何处理它。数据包不是为OS准备的，而是为在计算机上运行的某个进程准备的。同样，到达办公楼的包裹不是给传达室工作人员的，而是给办公楼里其他人的。操作系统检查端口号并把入站数据传递给监听指定端口的进程，就像传达室工作人员检查包裹收件人姓名或办公室号码把包裹送达正确收件人一样。

在0～1023范围内的网络端口被称为已知端口，而在1024～49151范围内的端口则可以在IANA（Internet Assigned Numbers Authority）中注册，称为注册端口。编码大于49151的端口是动态端口。从技术上来说，任何具有足够权限的进程都可以监听系统上还未使用的任意端口，这可能会忽略该端口号的典型用例。但是，当客户端应用程序想连接到另一台计算机上的远程服务时，它需要知道使用哪个端口，所以标准化端口号是有意义的。例如，Web服务器通常监听端口80和端口443（用于加密连接）。Web浏览器假设它应该使用端口80和端口443，除非另有指示。

练习11-2：研究常用端口

查找常用应用层协议的端口号。域名系统（Domain Name System，DNS）、安全壳（SSH）和简单邮件传输协议（Simple Mail Transfer Protocol，SMTP）的端口号是什么？你可以在网上搜索或者查看IANA注册表（http://www.iana.org/assignments/port-numbers）来获得这个信息。IANA列表有时会使用出乎预料的术语来表示服务名。例如，DNS只被简单地列为“domain”。

服务器使用众所周知的端口以方便客户端连接。然而，绝大多数网络通信是双向的（客户端发送请求，服务器响应），所以客户端也需要有一个开放端口，以便接收来自服务器的数据。客户端只需要暂时打开这样的端口，时间足够它完成与服务器之间的通信即可。这种端口被称为临时端口，由操作系统中的网络组件分配。例如，客户端Web浏览器通过端口80连接Web服务器，且客户端上的临时端口也是打开的，假设端口号为

61348。客户端把其Web请求发送到服务器上的端口80，服务器把响应发送给客户端上的端口61348。

IP地址加上端口号形成一个端点（endpoint），端点的一个实例称为套接字（socket）。套接字可以监听新连接，也可以表示已建立的连接。如果多个客户端连接到同一个端点，那么，每一个都有自己的套接字。

注意

请参阅设计31查看Raspberry Pi的端口使用情况。

11.2.4 应用层

应用层是互联网协议套件的最后一层，也是最高层。虽然较低的三层为互联网上的通信提供了通用基础，但应用层协议关注的是完成特定任务。例如，Web服务器使用超文本传输协议（HyperText Transfer Protocol，HTTP）检索和更新Web内容。邮件服务器使用简单邮件传输协议（SMTP）收发邮件消息。文件传输服务器使用文件传输协议（File Transfer Protocol，FTP）来做什么？传输文件！换句话说，应用层是我们获得描述应用程序行为的协议的地方，而这个协议栈的较低层是"管道"，使应用程序能在互联网上做它们想做的事情。完整的四层模型示意图如图11-12所示。

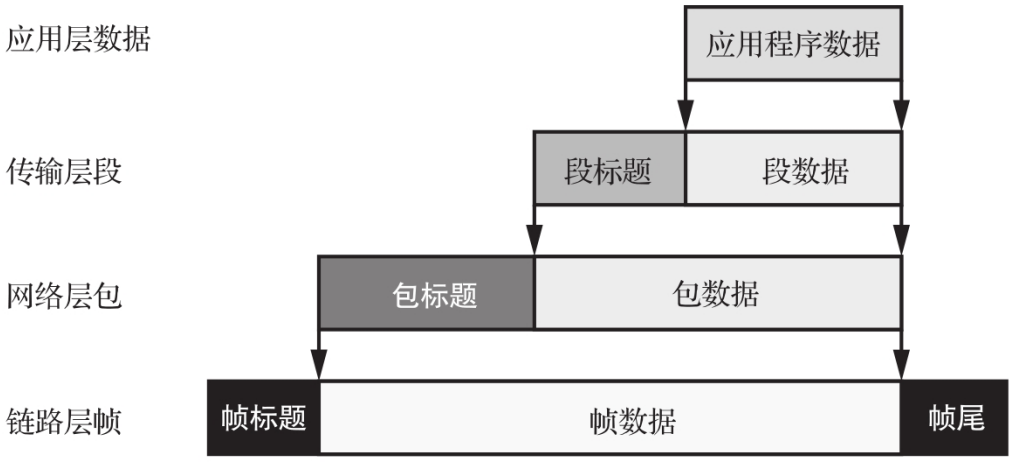

图11-12 应用层数据包含在段数据部分

图11-12是一个分层视图，它展示了每层是如何包含在下一层的数据负载中的。图11-13把所有的层组合在一起，我们可以看到发送给互联网上某个设备的一帧内容的表示形式。

图11-13 一个包含IP包、TCP段和应用程序数据的帧

我们从最接近硬件的层开始，自下而上地遍历了网络帧的内容。当一帧被主机接收时，主机会按相同的顺序（从链路层一直到应用层）来处理它。相反，当从主机发出一帧时，帧内容会按逆序组装。进程准备应用程序数据，该数据被包含到段中，再到包中，最后到帧。

11.3 游历互联网

现在你已经熟悉了四层TCP/IP网络模型中的每一层，接下来我们通过一个例子来看看数据是如何在互联网上传输的。我们将看到沿途各种设备是如何与每个网络层交互的。图11-14展示了左上角客户端与左下角服务器之间的通信。

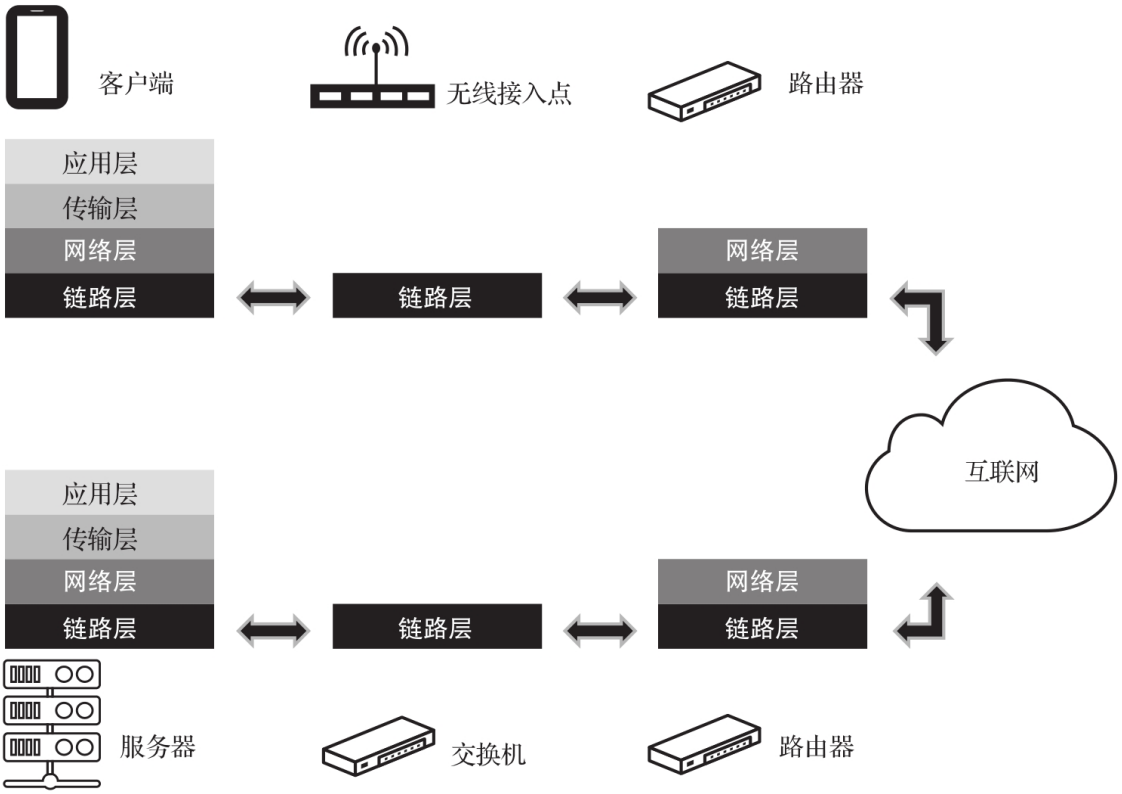

图11-14 不同设备在网络栈的不同层进行交互

我将建立图11-14中的场景。客户端设备（图中左上角）连接到一个无线Wi-Fi网络。这个网络通过路由器连接到互联网。在其他地方（图中左下角）有个服务器，它通过交换机和路由器建立了到互联网的有线连接。客户端设备的用户打开Web浏览器，请求在服务器上托管的网页。为了简单起见，假设客户端已经知道了服务器的IP地址。

客户端上的Web浏览器用HTTP"说话"，它是Web的应用层协议，所以它形成了针对目标服务器的HTTP请求。然后，浏览器把HTTP请求传递给操作系统的TCP/IP软件栈，要求把数据传送到服务器——明确服务器的IP地址和端口80（HTTP的标准端口）。接着，客户端操作系统上的TCP/IP软件栈把HTTP负载封装到TCP段（传输层）中，并在段标题将目标端口设置为80。如果有必要，TCP会把应用层数据分成多段，每段都有自己的段标题。客户端上的网络层软件再把TCP段封装进IP数据包中，IP数据包的包标题中包含了服务器的目标IP地址。同样，如果有必要的话，IP会把数据包分成较小的片段，为在网络链路中传送做好准备。在客户端的链路层上，IP包被

封装进帧中，帧标题中有本地路由器的MAC地址。这个帧由客户端设备的Wi-Fi硬件进行无线传输。

无线接入点接收该帧。这个接入点在链路层工作，它把帧发送给路由器。路由器检查网络层数据包以确定目标IP地址。为了到达服务器，请求需要途径互联网上的多个路由器。本地路由器把数据包封装到一个新的帧中——该帧带有新的目标MAC地址（下一个路由器的地址），然后发送这个新的帧。路由过程通过互联网上的多个路由器持续进行，直到请求到达服务器所连接的子网上的路由器。

最后一个路由器把数据包封装到适合服务器本地网络的帧中。这个帧的帧标题包含了服务器的MAC地址。该服务器子网上的交换机查看帧中的MAC地址，并把帧转发到合适的物理端口。交换机不需要查看更高的层。服务器接收该帧，网络接口驱动程序把TCP/IP包传递给TCP/IP软件栈，然后TCP/IP软件栈再把HTTP数据传递给监听TCP端口80的进程。Web服务器软件监听端口80，处理请求。这包括回复客户端，为此，整个过程会再次发生，只不过顺序相反。

注意

请参阅设计32查看从Raspberry Pi到互联网上某个主机的路由情况。

11.4 互联网基础功能

TCP/IP为数据在互联网上的可靠传输提供了必要的管道，而其他协议则提供了其他的物联网基础功能。这些功能作为应用层协议来实现。现在我们来看两个这样的协议（DHCP和DNS）以及一个转换IP地址的系统（NAT）。

11.4.1 动态主机配置协议

为了与其他主机通信，互联网上的每个主机都需要一个IP地址、子网掩码以及其路由器的IP地址（也被称为默认网关）。IP地址是如何分配的？

可以为设备提供一个静态IP地址，这需要有人编辑设备上的配置，并手动设置其IP信息。有时候，这是有用的，但它要求用户确保待分配IP地址还未被使用，且对子网是有效的。绝大多数最终用户不具备手动配置设备IP设置的专业知识，同时也不想处理手动配置的麻烦。幸运的是，大多数IP地址是用动态主机配置协议（Dynamic Host Configuration Protocol，DHCP）动态分配的。有了DHCP，当设备连接到网络时，它会收到一个IP地址和相关信息，无须用户干预。

要使DHCP在网络上可用，必须把网络上的一个设备配置成DHCP服务器。这个服务器有一个IP地址池，这些IP地址允许被分配给网络上的设备。DHCP会话如图11-15所示。

让我们来看看图11-15。当设备连接到网络时，它广播一条消息以发现DHCP服务器。广播是特殊类型的包，它发给本地网络上的所有主机。当DHCP服务器接收到这条广播后，它向客户端设备提供一个IP地址。如果客户端想接受提供的IP地址，它就以请求被提供地址的方式回复服务器。然后，DHCP服务器对这个请求进行应答，IP地址就此被分配给该客户端。IP地址是租给客户端的，如果客户端不续租，这个IP地址最终会过期。

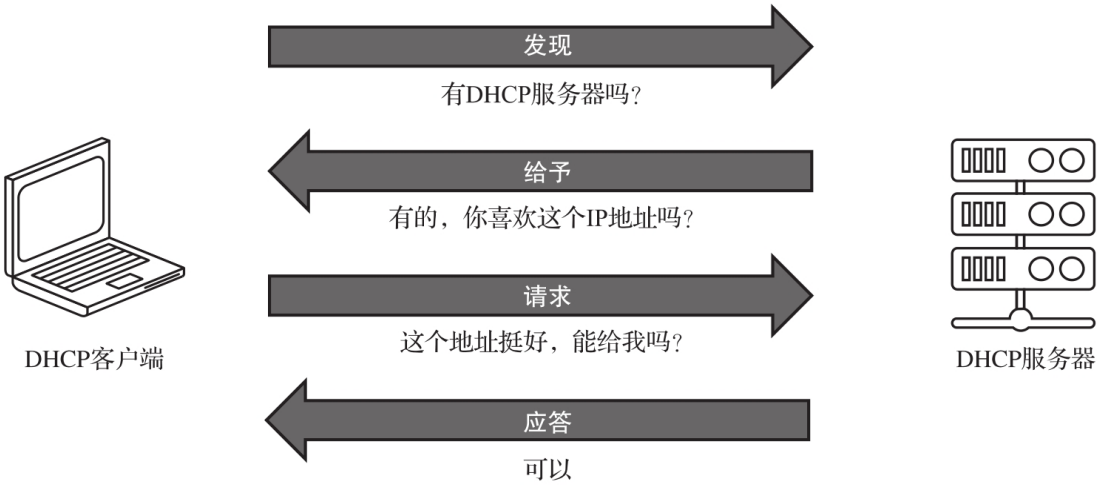

图11-15 DHCP会话

注意

请参阅设计33查看Raspberry Pi使用DHCP租到的IP地址。

11.4.2 私有IP地址和网络地址转换

可用的IP地址是有限的，所以大多数家庭互联网服务提供商（Internet Service Provider，ISP）只为一个客户分配一个IP地址。这个IP地址被分配给直接连接到ISP网络的那个设备（通常是一个路由器）。但是，很多客户的家庭网络上都有多个设备。让我们来看看如何利用私有IP地址和网络地址转换功能来让多个设备共享一个公网IP地址。

某些IP地址范围被视为私有IP地址，这些地址被用于私有网络，比如那些家里或办公室里的网络，其中的设备不直接连接到互联网。任何能匹配10.*x*.*x*.*x*、172.16.*x*.*x*或192.168.*x*.*x*模式的地址都是私有IP地址。任何人不经允许就可以使用这些范围内的IP地址。问题是私有IP地址是不可路由的——它们不能在公网上使用。家庭网络上的DHCP服务器可以分配这些地址，而不用担心其他网络是否会使用相同的地址。和公网IP地址必须唯一不同，私有IP地址旨在在多个私有网络上同时使用。多个网络是否使用相同的地址并不重要，因为这些地址无论如何都不会在私有网络之外被看到。私有IP地址解决了ISP只能为家庭或企业提供单个公网IP地址的问题，但是，如果私有IP地址不能在互联网上路由，那么它们又有什么用呢?

网络地址转换（Network Address Translation，NAT）允许私有网络（通常是家庭网络）上的所有设备都使用互联网上相同的公网IP地址。当数据包经过NAT路由器时，该路由器修改这些包中的IP地址信息。当来自私有家庭网络的数据包到达NAT路由器时，其源IP地址字段会被修改为公网IP地址，如图11-16所示。