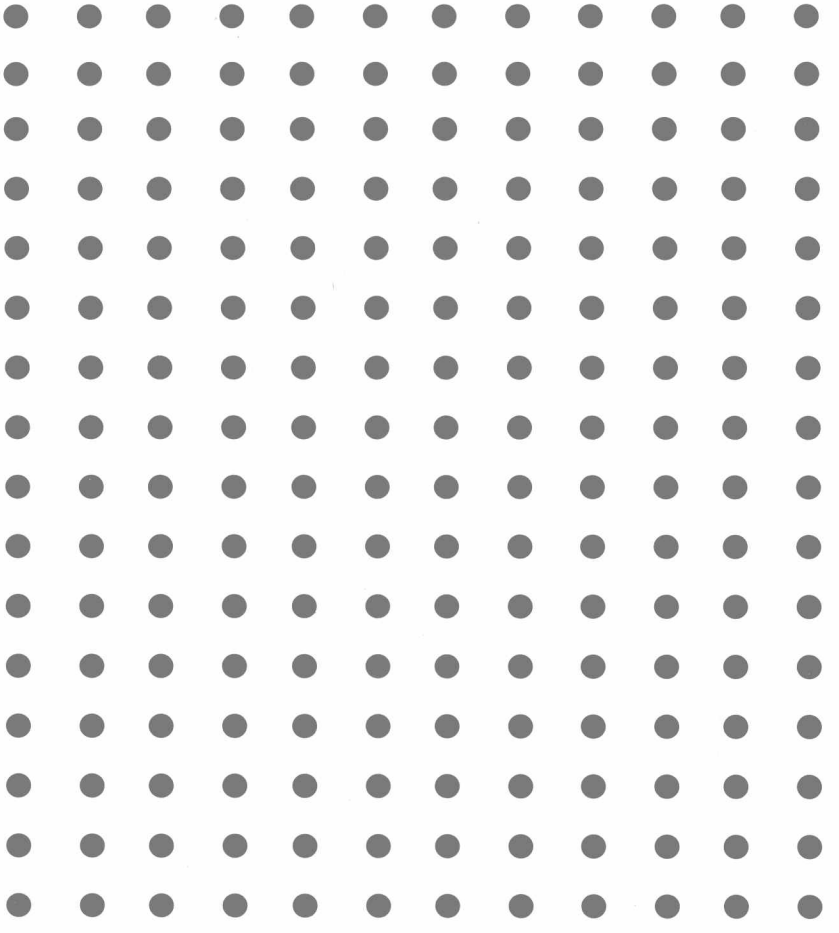
徐宇杰 编著

TCP/IP协议深入分析

清华大学出版社

TCP/IP协议深入分析

徐宇杰 编著



清华大学出版社
北京

内 容 简 介

本书详细说明 TCP/IP 协议簇，以截屏和协议包结构为手段，介绍了 TCP/IP 各层的细节，对 IP、TCP、UDP、ARP、ICMP、HTTP、Telnet、FTP 与 TFTP、POP3 与 SMTP、DHCP 协议进行了深入剖析。本书实例丰富，图文并茂，注重理论与实践结合，降低了读者的学习难度，激发了读者学习兴趣和动手欲望。

本书可作为网络从业人员的专业学习和参考用书，也可作为大中专院校网络课程教材。

图书在版编目(CIP)数据

TCP/IP 协议深入分析/徐宇杰编著. —北京：清华大学出版社，2009.2
ISBN 978-7-302-18416-4

Ⅰ. T…　Ⅱ. 徐…　Ⅲ. 计算机网络—通信协议　Ⅳ. TN915.04

中国版本图书馆 CIP 数据核字(2008)第 125636 号

责任编辑：丁　岭　赵晓宁
责任校对：焦丽丽
责任印制：杨　艳
出版发行：清华大学出版社　　地　　址：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　邮　　编：100084
社　总　机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京市清华园胶印厂
经　　销：全国新华书店
开　　本：185×260　印　张：10.25　字　数：250 千字
版　　次：2009 年 2 月第 1 版　　印　　次：2009 年 2 月第1 次印刷
印　　数：1～3000
定　　价：19.00 元

FOREWORD

前言

Internet不仅深刻地改变了整个IT行业的格局和计算模式，也深刻改变了经商的方式和人们的生活方式，网络已经成为整个基础设施中的重要部分。

Internet也叫网际网，也就是网络和网络互联。因此，当前的网络技术不仅仅是Socket编程这样点对点的通信技术和Windows Server下的Web服务器配置这样的系统管理技术。从TCP/IP的视角来看，所谓网络技术实际上包括计算机网络原理、计算机网络设计、计算机网络工程、计算机网络协议、计算机网络互联、计算机网络应用等几个方面的范畴。一个网络的互联是跨越不同网络层次的一个过程。因此，基于互联网的网络技术的设计、实现、管理和排错，需要自底向上的多层次知识。

作者从20世纪90年代初期就开始学习和从事TCP/IP网络技术，工作涉及局域网、园区网、城域网、跨区域网的设计、实现和管理的全过程，经历了从局域网到互联网那激动人心的变化。又有多年大型网络服务器的管理经验，积累了丰富的网络技术实践经验。作者多年的学习、管理、开发和应用，深感目前计算机网络技术方面的书籍往往只介绍一个方面的技术，顾此失彼的多，导致多层结构的网络技术被支离破碎地介绍，很多技术方面的介绍流于表面，学习者很难透彻了解。理论不深刻，实践不实际的情况非常普遍。同时，很多网络厂商又从各自的市场利益出发，积极推广相关的认证和课程。这些课程对网络技术的应用起到了非常正面的作用。但是，这些课程指导书由于更偏向认证的各个知识点，实验内容也围绕认证服务，对于网络技术的细节缺乏深入分析。

如何兼顾理论和实际应用，一直是困扰计算机网络教学的一个难题。作者多年的实践经验一再证明没有扎实的网络理论知识，根本无法从事网络技术。另一方面，学习者只有通过真实项目的全过程实践才能真正掌握网络理论知识。

笔者5年前开始总结自己学习和实践经验，试图将实际项目中最常用、也是最核心的知识点加以梳理。此系列书籍的选题、内容选择、实验的准备

和测试，可以说是去粗取精的结果。本系列丛书涵盖了 TCP/IP 协议的分析、网络交换、网络路由、下一代互联网技术。有详细的包结构截屏和深入的细节分析，从基本理论学习和分析开始，逐步分析网络技术的各层细节，最后提供一个从布线开始，包含设计和配置、运行一个真实园区网络的综合实例教程。本系列丛书实例丰富、图文并茂，边讲解边操作，将网络技术的细节一一展现，降低了读者的学习难度，激发了学习兴趣和动手欲望，适合网络从业人员的专业学习和参考以及各个院校作为计算机网络的实例教学。

5 年的时间，对于网络技术来说似乎是一个漫长的过程，实际上是一个让时间来检验的优选过程。每本书经过多次修改，不仅进行理论内容的修正，更多的是实验的更新，甚至全部改写，试图在出版前反映最新的技术变化。但有可能百密一疏，望读者指正。

编　者

2008 年 6 月

CONTENTS

目录

CHAPTER 1

TCP/IP 协议概述 第 1 章

TCP/IP 是一个四层协议系统，TCP/IP 协议族是一组不同的协议组合在一起构成的协议族。如表 1-1 所示，每一层负责不同的功能。

表 1-1

TCP/IP	主要协议	主要功能
应用层	Http、Telnet、FTP、E-mail 等	负责把数据传输到传输层或者接收从传输层返回的数据
传输层	TCP、UDP	为两台主机上的应用程序提供端到端的通信。TCP 为两台主机提供高可靠性的数据通信，它所做的工作包括把应用程序交给它的数据分成大小合适的数据块交给下面的网络层，确认接收到的分组等。UDP 则为应用层提供不可靠的数据通信，它只是把数据包的分组从一台主机发送到另一台主机，但是不保证该数据能到达另一端
网络层	ICMP、IP、IGMP	主要为数据包选择路由，其中 IP 是 TCP/IP 协议族中最为核心的协议，所有的 TCP、UDP、ICMP、IGMP 数据都以 IP 数据包格式传输
链路层	ARP、RARP 和设备驱动程序及接口	发送时将 IP 包作为帧发送，接收时把收到的位组装成帧，同时提供链路管理，错误检测等

TCP/IP 协议族中 TCP 和 IP 只是其中的两种协议，其中 TCP 和 UDP 是两种最为著名的传输层协议，IP 是网络层协议。IP 和 TCP 这两个协议的功能不尽相同，它们是在同一时期作为一个协议来设计的，并且在功能上也是互补的，虽然它们可以分开单独使用，但是只有两者的结合，才能保证 Internet 在复杂的环境下正常运行。要连接到 Internet 的计算机，都必须同时安装和使用这两个协议，因此在实际中常把这两个协议统称作 TCP/IP 协议。

在 TCP/IP 协议族中，各层的关系如图 1-1 所示。

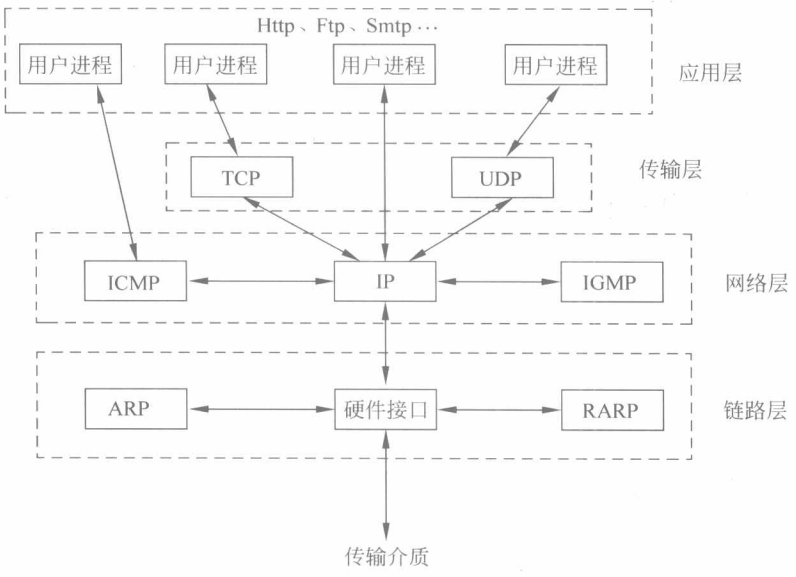


图 1-1

因特网控制消息协议(Internet Control Message Protocol,ICMP)是 IP 协议的附属协议。IP 层用它来与其他主机或路由器交换错误报文和其他重要信息。IGMP 是 Internet 组管理协议,它用来把一个 UDP 数据包多播到多个主机。

ARP(地址解析协议)和 RARP(逆地址解析协议)是某些网络接口(如以太网和令牌环网)使用的特殊协议,它们用来转换网络接口的物理地址和对应的 IP 地址。

当目的主机收到一个以太网数据帧时,数据就开始从协议栈的底部往上升,同时去掉各层协议封装的报文首部。每层协议盒都要去检查报文首部中的协议标识,以确定接收数据的上层协议。这个过程称作分用(Demult IP Lexing),如图 1-2 所示。

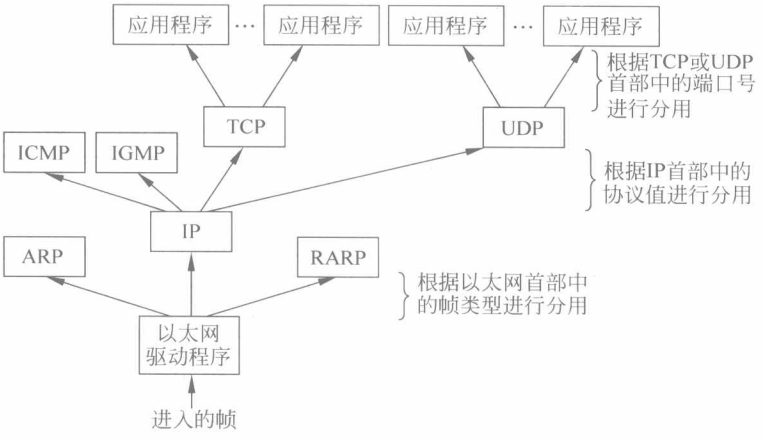


图 1-2

图 1-1 和图 1-2 中的协议分层并不是绝对的,拿 ICMP 和 IGMP 来说,在图 1-1 中,把它们与 IP 放在同一层上,那是因为事实上它们是 IP 的附属协议。但是在图 1-2 中,又把它们

放在 IP 层的上面，这是因为 ICMP 和 IGMP 报文都被封装在 IP 数据包中。

再比如 ARP 和 RARP，在图 1-1 中，把 ARP 作为以太网设备驱动程序的一部分，放在 IP 层的下面。在图 1-2 中，把它们放在以太网设备驱动程序的上方，这是因为它们和 IP 数据包一样，都有各自的以太网数据帧类型。

这里用两种图只想说明各种协议之间彼此的关系，它们之间并不是彼此孤立的。

1. IP 层

在 TCP/IP 协议族中，网络层 IP 提供的是一种不可靠的服务，它只是尽可能快地把数据从源结点送到目的结点，并不提供任何可靠性保证。在通信中，IP 层只负责数据的路由与传输，并不处理数据包的内容。例如 ICMP，TCP 或 UDP，这些协议是依赖 IP 层的传输功能来传送数据的。在通信双方的主机中，收到这些协议的数据包后，一般在通信的对应主机上，会有程序来处理这些数据。

2. TCP 层

TCP 层位于 IP 的上层，应用程序在 IP 网络上相互之间传输的标准传输协议有两个，一个是传输控制协议(TCP)，TCP 是目前 Internet 上使用的最重要的协议，它提供的是可靠的、可控制的传输服务，大部分 Internet 应用程序都使用 TCP，因为它的嵌入可靠性和流控制服务可确保数据不会丢失和被破坏。另一个是用户数据包协议(UDP)，它提供的服务轻便但不可靠。

IP 层提供了一种不可靠的服务，TCP 在不可靠的 IP 层上提供了一个可靠的传输层，TCP 采用了超时重传、发送和接收端到端的数据确认等机制来保证这种服务的可靠性。由此可见，传输层和网络层分别负责不同的功能。

CHAPTER 2

第2章 IP协议

IP是TCP/IP协议族中最为核心的协议。所有的TCP、UDP、ICMP及IGMP数据都以IP数据包格式传输。目前的协议版本号是4，因此IP有时也称作IPv4。未来的版本就是IPv6。

IP层只负责数据的路由与传输，并不处理数据包的内容。IP提供不可靠、无连接的数据包传送服务。

不可靠(unreliable)的意思是它不能保证IP数据包能成功地到达目的地。IP仅提供最好的传输服务。如果在传输过程中发生某些错误时，IP有一个简单的处理错误算法：丢弃该数据包，然后发送ICMP消息报给源发送端。

每经过一个系统(主机或路由器)都会检查这个包，如果这个包已经损坏或者经历过其他形式的暂时性失败，该包就会在此立刻被删除。

如果出现的问题是半永久性(semi-permanent)的，那么IP就会发送ICMP来给源发送者返回一个错误消息，将这次失败告诉源发送者，以让发送者修改引起失败的情况，然后删除该数据包。

暂时性的失败和半永久性失败之间的界限很重要。暂时性错误是指不是因为发送者的错误引起的(例如生命期超时发生的错误或者校验和计算错误)。半永久性失败是指包或网络出现了问题导致这条路径不能发送数据，这时最好把问题告诉发送者以便纠正错误。

IP提供无连接服务也就是IP并不维护任何关于后续数据包的状态信息。IP网络把每个IP数据包都当作一个完全独立的实体，每一个实体都通过当时最合适的路径进行传输。例如，如果一个用户想从一个远程Web服务器获取一个文档，为了返回请求的材料，这个服务器可能需要创建几个IP数据包。每个数据包都通过沿途的路由器来寻找最合适的路径进行传输。比如Web服务器往请求客户端发送的第一个数据包可能通过地下光纤电缆传输，而第二个可能通过卫星连接来传输，第三个可能通过传统的网络来传输。每个数据包之间是独立的，它的好处就是数据包不用按顺序到达目的地，因为某个数据包可能经过一个快速网络传输，而其他数据

包可能经过一个慢速网络传输，还有数据包可能会重复，导致某个包有多个拷贝到达目的地。

另外，网络把每个数据包都看作是单个的实体，它自身不用负责跟踪所有的连接，这样网络设备可以专注于传输数据包，这个特性使得整个性能能够提高到硬件允许的水平，而对内存和 CPU 要求却尽可能地低。

由于每个数据包的处理是相互独立的，这样当数据包到达目的地时就会失序，怎么样来组织起这些失序的数据包呢？原来在 IP 首部中包含着一些信息，以让接收端能正确组装这些数据包。

2.1　IP 分片和重组

IP 分片与网络的 MTU 有关。先来看看 MTU：

MTU(Maximum Transmission Unit)是网络上传送的最大数据包。MTU 的单位是字节。例如，Ethernet 在一个帧中仅仅能够发送 1500 个字节，而 16MB/s Token Ring 的典型的 MTU 每个帧是 17 914 个字节。

RFC791 规定，MTU 最大是 65 535 个字节，最小是 68 个字节。当 IP 层接收到一份要发送的 IP 数据包时，如果数据包大于发送系统的本地 MTU 的话，这时系统就得把数据包分成多个片(fragmentation)来发送。在数据包传输过程中，如果 IP 数据包太大而不能通过发送者和最终接收者之间的某个网段时，路由器也会把包分段，使得数据包能顺利通过这个网络。

分片后的 IP 数据包，只有到达目的地才进行重新组装。重新组装由目的端的 IP 层来完成，其目的是使分片和重新组装过程对传输层(TCP 和 UDP)是透明的。已经分片过的数据包有可能会再次进行分片(可能不止一次)。

当 IP 数据包被分片后，每一片都成为一个分组，具有自己的 IP 首部，这些分片后的数据包相互独立，会选择各自的最佳路由到达目的地。这样，当数据包的这些片到达目的端时有可能会失序，接收端凭件在 IP 首部中的信息正确组装这些数据包片。

虽然数据包可以分片，但是要尽量避免分片，因为 IP 层本身没有超时重传的机制，由更高层来负责超时和重传(TCP 有超时和重传机制)。当某片报文丢失后，TCP 在超时后会重发整个 TCP 报文段，而不是重传数据包文段中的一个数据包片。分片还会增加丢包率，降低网络速度。

2.2　分片规则

分片仅仅出现在数据包的数据部分。

分片过程不包括数据包的首部。如果源数据包是 4464 字节，那么这个数据包中至少有 20 字节是用来储存首部信息，这意味着数据部分是 4444 字节。要分片的就是这 4444 字节。

每一个分片都会产生一个包含它自己 IP 首部的新包，这个 IP 首部至少耗用新包中的 20 字节。

IP 分片必须以 8 字节的倍数分片。如一个数据包包含 256 字节的数据，但前一个片段仅能容纳 250 字节，那么第一个片段仅仅包含 248 字节(248 是小于 250 的可以被 8 整除的整数)。剩下的 8 字节就在下一个片段发送。

要尽量避免 IP 分片，TCP 的 Path MTU Discovery 机制可以提供一个最优值来避免数据分片。

2.3 IP 数据包结构

IP 数据包包含要发送的数据和相关的 IP 首部，如图 2-1 所示，这是个 IP 数据包的结构，显示了一个 IP 数据包所包含的信息。

图 2-1

IP 数据包中每个字段的意义如表 2-1 所示。

表 2-1

字　段	位数	用 法 说 明
版本(Version)	4	说明用以创建该数据包的 IP 版本。所有接触该数据包的设备都必须支持本字段显示的版本。大部分 TCP/IP 产品都使用 IPv4
首部长度(Header Length)	4	以 32 位为单位表明 IP 首部的长度。因为几乎所有的 IP 首部都是 20 字节长，这个字段的值几乎总是 5
服务类型(Type-of-Service Flags)	8	给应用程序、主机和 Internet 上的路由器提供一个优先级服务。在这个字段设置合适的标志，应用程序可以要求这个数据包得到高优先级，而让其他数据包等待
总长度(Total Packet Length)	16	以字节为单位说明全部 IP 包的长度，包括首部和主体部分
标识(Fragment Identifier)	16	标识数据包，在出现分片并想把片段合并成原状时是有用的
标志(Fragmentation Flags)	3	说明可能出现的任何分片的某些方面，也提供了分片控制服务，例如不让路由器分片某个包
偏移(Fragmentation Offset)	13	说明这个片段提供的源 IP 数据包的字节范围，用 8 字节的偏移表示

续表

字　段	位数	用 法 说 明
生存时间(Time-to-Live)	8	说明数据包在不可发送和破坏之前还可以经过的跳数
协议(Protocol Identifier)	8	说明储存在 IP 数据包主体的高层协议
首部校验和(Header Checksum)	16	用来储存 IP 首部的校验和
源 IP 地址(Source IP Address)	32	用来储存最初发送该数据包的主机的 32 位的 IP 地址
目的 IP 地址(Destination IP Address)	32	用来储存该数据包到达目的系统的 32 位的 IP 地址
选项(options)	可变	就像 IP 用 type-of service 标志提供了一些优先级服务一样，附加的特殊处理选项能够通过 Options 字段定义。这些选项包括 source routing，timestamp 和其他一些选项。这些选项很少用，这也是会导致 IP 首部长度超过 20 字节的唯一原因(这个选项是可选择的)
Padding(如果需要的话)	可变	IP 数据包的首部的长度必须是 32 的倍数。如果首部中引入了某些选项，首部必须填充到能够被 32 整除的位数
数据(data)	可变	IP 数据包的数据部分。正常情况下，会包含一个完全的 TCP 或 UDP 信息，但是它也可以是其他 IP 数据包的一个片段

下面通过一个实际的 FTP 数据包来分析一下 IP 数据包的结构。

1. Version 和 Header Length

如图 2-2 和图 2-3 所示，可以看到 IP 数据包前面两个字段 Version 和 Header Length。

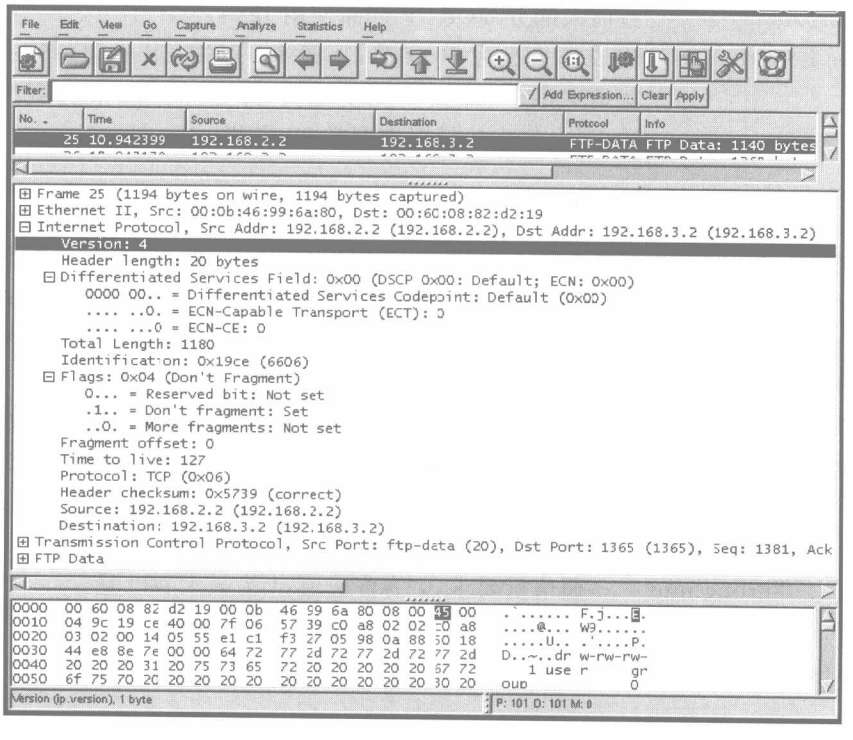


图　2-2

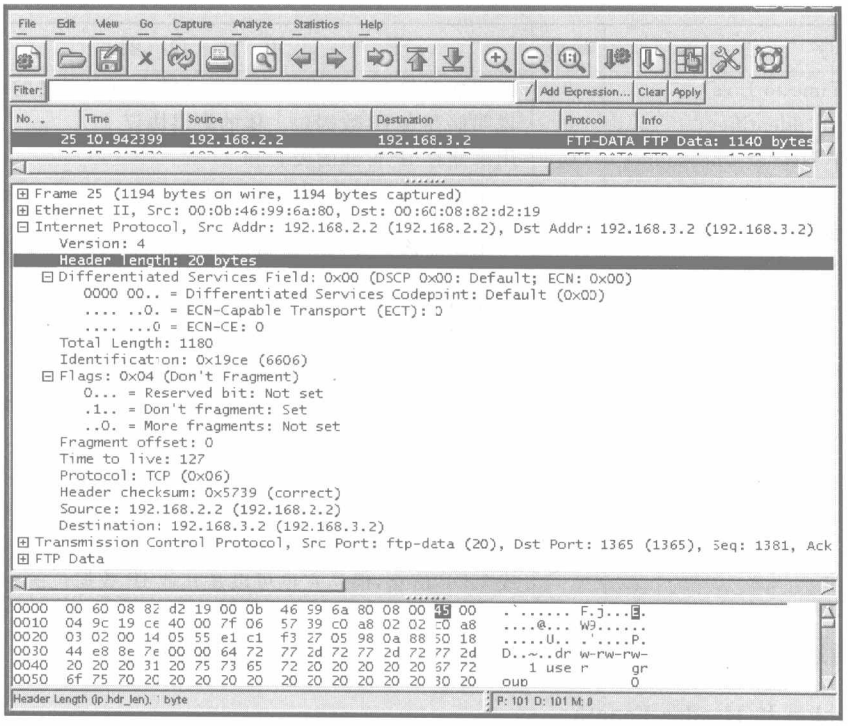


图 2-3

Version：4　　目前大部分 TCP/IP 产品使用 IPv4。

Header Length：20　　IP 首部大小为 20 字节，从图 2-1 可知，IP 封包每行有 32 位，也就是 4 字节，那么 5 列就是 20 个字节。如果选项没有设定的话，那么包头的最短长度是 20 字节。

2. Differentiated Services Field

接下来的字段 Differentiated Services Field，如图 2-4 所示，它是为 IP 数据包提供优先级能力，这些优先级能力作用于能够利用它们的应用程序、主机和路由器上。通过适当的设置这些字段，一个应用程序可以请求它产生的数据包得到优先于其他等待通过的数据包的服务。

尽管该标志自 IPv4 首次发布以来就可以用了，但真正使用它的却只是少数应用程序。此外，只有少量的 IP 软件包和路由器支持它，这使得应用程序使用它没什么意义。

000..... Routine　　优先权要求，设为 0 为普通优先级，否则，数值越高越优先。

...0.... Delay　　延迟要求，0 是普通值，1 为最小延迟。

....0... Throughput　　通信量要求，0 为普通值，1 为最大吞吐量。

.....0.. Reliability　　可靠性要求，0 为普通值，1 为最高可靠性。

.....00 Not Used　　未使用。

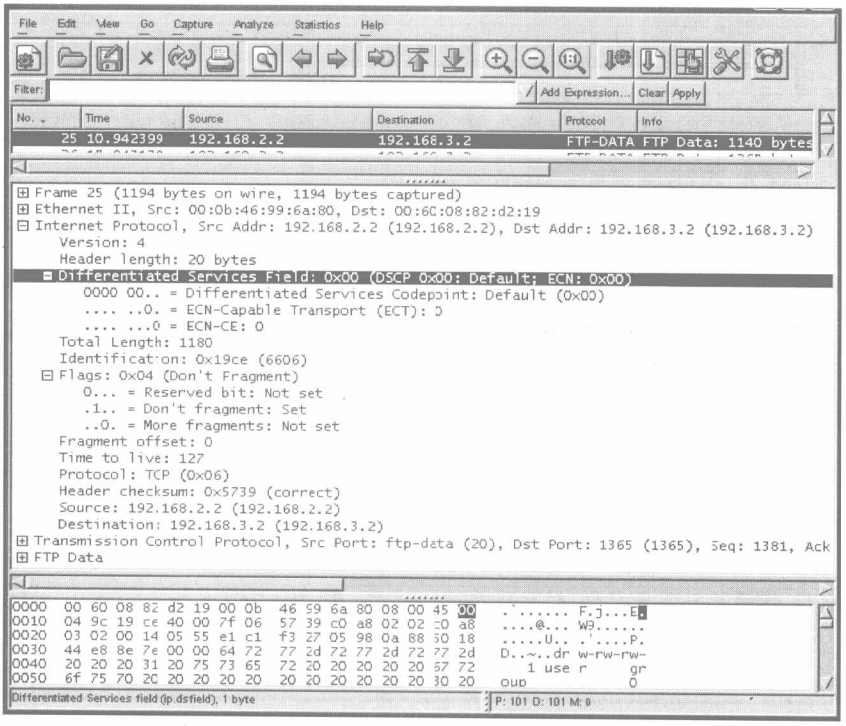


图　2-4

3. Total Length

说明整个 IP 数据包的大小，包括首部和数据部分，以字节表示。如图 2-5 所示，可以看到 Total Length：1180，同时注意到图最下方有个黑色加强的部分 04 9c，这个数字是 1180 的十六进制表示。

这个字段的主要目的是告诉系统该包的终止位置：total packet length－(减去) header length。

4. Identification

如图 2-6、图 2-7 所示，标识字段唯一地标识主机发送的每一份数据包。产生的每一个数据包都有 16 位的序列号，用来让发送系统和接收系统识别该数据包。通常每发送一份报文它的值就会加 1。

当要发送一个数据包分片的时候，会把这个字段的内容复制到每个片中，表示这些被分割的片属于同一个数据包。

图 2-6 和图 2-7 中两个数据包的序号分别为 6606 和 6607，在图 2-7 中，数据包的 Identification 字段的值是图 2-6 中的值加 1。

5. Flags

如图 2-8 所示，这个字段占有 3 位，它们的表示的意义如下：

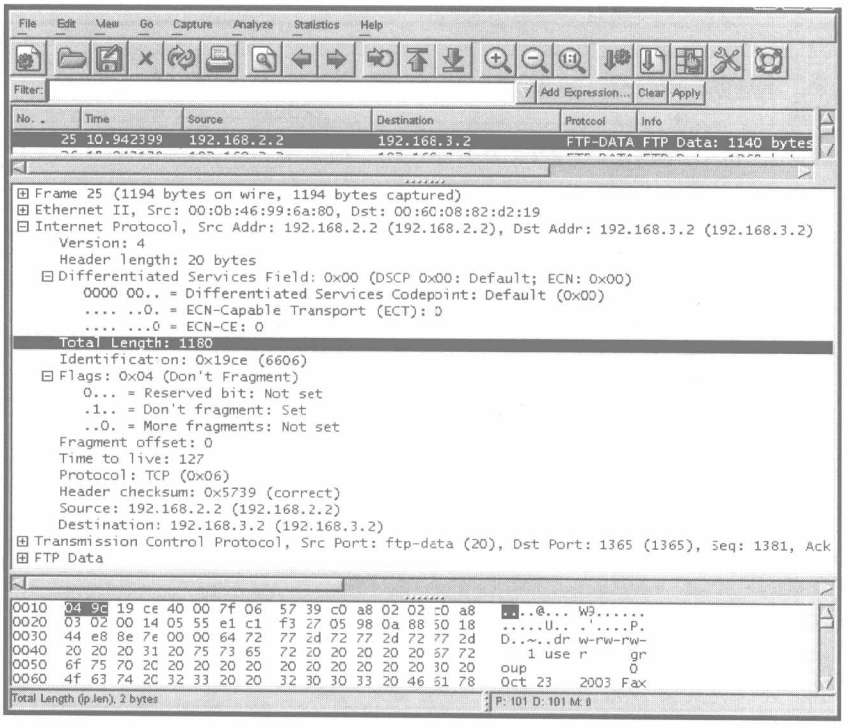


图 2-5

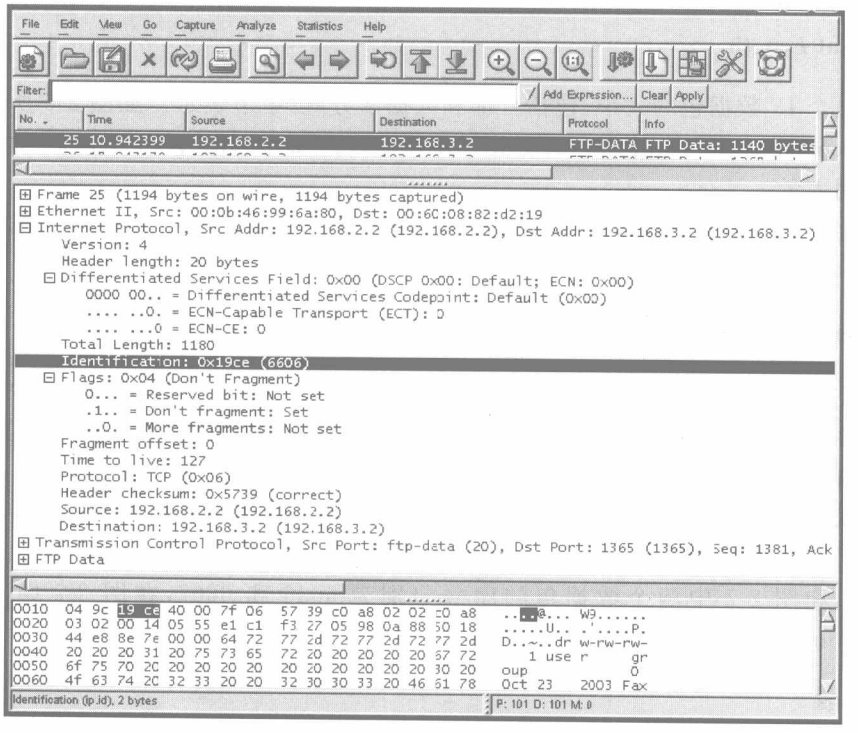


图 2-6