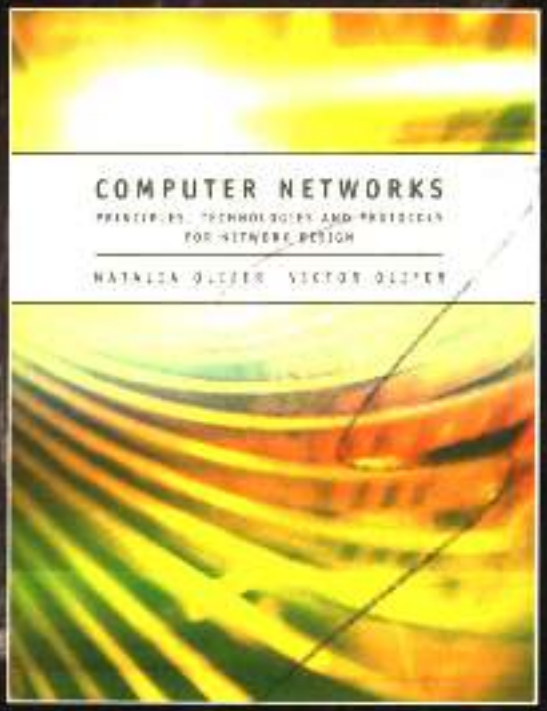
计 算 机 科 学 丛 书

计算机网络

网络设计的原理、技术和协议

（俄）Natalia Olifer （乌）Victor Olifer 著 高传善 等译



Computer Networks

Principles, Technologies and Protocols for Network Design

机械工业出版社
China Machine Press

计算机网络 网络设计的原理、技术和协议

现代复杂的计算机网络由众多层次、架构和协议组成，要想清晰地勾勒其整体画面非常困难。网络的关键组成部件不仅仅需要被孤立考察，作为异构系统的一部分，每部分还需要与各种各样的网络技术联合工作。

本书广泛深入地讲解了一系列复杂的课题，并涵盖组网技术的基础理论以及组网中所遇问题的实际解决方法。基于现代的集成环境，作者的讲解方法可以帮助读者将网络理解为一个整体，而不仅仅是分散的部件的组合。

本书适合大学本科生、研究生以及IT专业人员阅读和使用。通过本书，读者可以掌握网络原理的基础知识，理解局域网(LAN)和广域网(WAN)传统的和最新的技术与特性，以及学习设计、管理主流和复杂网络的方法。本书还讨论了对以下问题的基本解决方法：数据编码、差错检测、媒体访问、路由、流量和拥塞控制以及端到端的运输。

本书提供了教学支持网站可供授课教师获取与本书相关的补充材料，其中包括组网问题的描述及其解决方法。

WILEY
www.wiley.com

上架指导：计算机/计算机网络

ISBN 978-7-111-22885-1

投稿热线：(010) 88379604
购书热线：(010) 68995259, 68995264
读者信箱：hzjsj@hzbook.com

华章网站 http://www.hzbook.com

网上购书 www.china-pub.com

ISBN 978-7-111-22885-1
定价：68.00 元

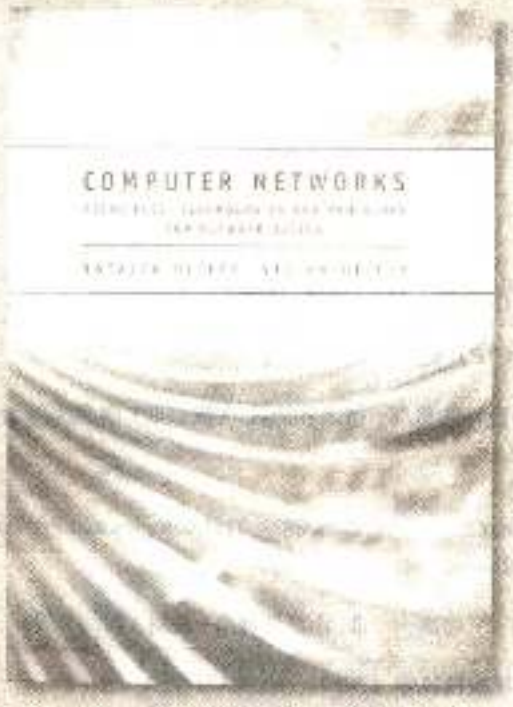
TP393/556
2008

计 算 机 书

计算机网络

网络设计的原理、技术和协议

(俄) Natalia Olifer (乌) Victor Olifer 著 高传善 等译



Computer Networks

Principles, Technologies and Protocols for Network Design

机械工业出版社
China Machine Press

本书是计算机网络的一本基础教材，不仅涵盖了计算机网络的主要基础知识和最新技术内容，还对各种网络技术的细节和使用设备的特性做了综合的介绍和分析。全书共分为五大部分，总计24章。主要内容包括：网络基础、物理层技术、局域网、TCP/IP网际互联、广域网。本书还包含了有关Cisco认证考试所需要的部分理论知识。

本书可供需要掌握计算机网络基础理论和实践知识的本科生和研究生作为教材或参考书。也可供网络专业人员和IT专业人士使用。

Natalia Olifer, Victor Olifer: Computer Networks: Principles, Technologies and Protocols for Network Design (ISBN: 0-470-86982-8)

Authorized translation from the English language edition published by John Wiley & Sons, Inc.

Copyright©2006 by John Wiley & Sons, Inc.

All rights reserved.

本书中文简体字版由约翰-威利父子公司授权机械工业出版社独家出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。

版权所有，侵权必究。

本书法律顾问　北京市展达律师事务所

本书版权登记号：图字：01-2006-2850

图书在版编目（CIP）数据

计算机网络：网络设计的原理、技术和协议/（俄罗斯）奥里夫等著；高传善等译. -北京：机械工业出版社，2008.1

（计算机科学丛书）

书名原文：Computer Networks: Principles, Technologies and Protocols for Network Design

ISBN 978-7-111-22885-1

Ⅰ.计…　Ⅱ.①奥…　②高…　Ⅲ.计算机网络-教材　Ⅳ.TP393

中国版本图书馆CIP数据核字（2007）第185352号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）

责任编辑：王　璐

北京市慧美印刷有限公司印刷　·　新华书店北京发行所发行

2008年1月第1版第1次印刷

184mm×260mm　·　35.25印张

定价：68.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线：（010）68326294

译者序

本书译自Natalia Olifer和Victor Olifer所著《Computer Networks : Principles, Technologies and Protocols for Network Design》(John Wiley & Sons, Ltd, 2006)。本书拥有庞大的读者群，最初是用俄文写的，已出版了三个版本，2006年又从俄文翻译成英文出版。本书内容丰富，共有24章，作者按照网络基础、物理层技术、局域网、TCP/IP网际互联和广域网五大模块进行讲解，既涵盖网络技术的基础理论，又通过案例的研究给出了组网中所遇问题的实际解决方法。全书文字深入浅出，在编排上有较大更新，并注意了论题的取舍和平衡，因而对网络的概念讲解清楚，并着重讲解了各部件间如何通过各种技术一起工作而构成一个复杂的计算机网络，有助于读者对网络有一个整体的了解。从书中的内容不难看出，作者具有讲课和实际工作两方面的丰富经验。

本书每章后面都有复习题与练习题，复习题可以帮助读者复习每章的重要内容，练习题则着重强化读者学会运用所学的知识去解决实际的问题的能力。本书最后分别列出了参考文献与推荐阅读的书。本书还为教师提供了一个支持网站www.wiley.com/go/olifer，采用本书作为教材的教师可以登录该网站并下载课件等教辅资料。如作者在前言中所说，本书内容已经在各类听众中被成功检验。这些听众来自不同的群体并具有不同的专业兴趣，其中包括大学本科生和研究生、IT部门的负责人、网络管理员和集成商等。因而，虽然该书主要面向那些希望能够掌握系统的网络理论和实践知识的本科生和研究生，但是对有经验的网络专业人员也同样具有参考价值——他们可以在本书中学到过去在实际工作过程中未曾遇到的新技术，或者借助书中的内容对自身已有的知识进行整理。本书也可作为实际工作的参考书，在书中可以找到某些特定协议或帧格式的详细描述。此外，本书还提供了准备Cisco认证考试（如CCNA、CCNP、CCDP和CCIP）所需要的部分理论知识。

本书由复旦大学高传善主持翻译和统稿工作。同时，高传善还翻译了前言、结束语和封底。李莹翻译了第1~7章，王志伟翻译了第8~11章，胡磊翻译了第12~14章，祝轶群翻译了第15、16章，李冰峰翻译了第17~21章，张华翻译了第22~24章，最后由高传善统稿与审校。在翻译过程中我们力求忠实于原著，但限于时间及水平有限，不当之处在所难免，欢迎读者批评指正。

译　者

复旦大学计算机科学与工程系

二〇〇七年夏

译者简介



高传善，1942年生，1963年毕业于复旦大学，1981-1983年在美国伊利诺大学（UIUC）计算机科学系作访问学者。现为复旦大学计算机科学与工程系教授、博士生导师和计算机与通信实验室主任，并兼任教育部全国计算机等级考试（NCRE）委员会委员、福建省人民政府顾问团顾问。

译者长期从事计算机系统、软件和应用方面的教学与科研工作，业务专长为数据通信、计算机网络、分布式系统及应用，在国内外学术刊物和会议上发表论文180余篇，正式出版著译作30余本。科技成果曾获省部级科技进步特等或一等奖3项、二等奖2项、三等奖3项。1992年开始享受国务院政府特殊津贴。1993年获光华科技基金奖二等奖。1995年被评为上海市优秀教育工作者。1999年获上海市人民政府决策咨询研究成果三等奖。2002年《数据通信与计算机网络》获全国普通高等学校优秀教材一等奖。2004年“计算机网络实践课与教材建设”获上海市优秀教学成果二等奖。2006年被授予“复旦大学教学名师”称号。迄今已培养硕士研究生78名、博士研究生27名、博士后3名和工程硕士研究生19名，其中已获硕士学位63名、已获博士学位17名、已出站博士后3名和已获工程硕士学位13名。

前　言

入门

本书是计算机网络基础教材，覆盖了这门快速发展的学科领域内的主要论点、问题和技术，对各种网络技术的细节和所用设备的特性作了综合的介绍和分析。本书的内容是作者在各大学、商业培训中心和大型企业培训中心多年教学经验积累的成果。

下面是本书的一些特点：

- **重点研究网络运输功能**。运输功能是确保计算机间数据传输的功能，继而形成计算机网络。本文将重点研究网络体系结构、电信设备的主要运行原理，以及网络用来运输数据的主要协议，包括网际协议（IP）、以太网、蓝牙、IEEE 802.11（Wi-Fi）、异步传送模式（ATM）和帧中继等。对于各种网络服务，我们重点讨论直接支持网络运输功能的服务（比如域名系统、动态主机配置协议、VPN和IPSec等），而不是为计算机用户提供的服务（如Web服务）。
- **不限于IP**。因特网的成功使IP成为构建互联网络的主要方式。但是，本书不仅仅局限于将不同种类的网络互联为一个统一的超级网络（如因特网）的IP技术，而对以太网和ATM这些组网技术也进行了详细讨论，这也是构建统一网络的基础。这两类网络技术对组建有效运行的现代网络是同等重要的，但这一点往往由于当前人们更偏爱IP（很明显的“追随潮流效应”）而被破坏，本书要力图恢复这种平衡。
- **结合计算机科学和计算机工程方法**。读者在书中会发现有关电信网络运行原理和通信协议操作算法的描述。这些知识通常被认为属于计算机科学体系，它对推动研究工作的成功是必要的。此外，本书还提供了通信设备的大量技术细节和设计各种类型网络的实际例子。这在你准备工程工作时是很有用的，而工程工作对于任何电信专业人员都是非常重要的。
- **各种类型电信网络的融合**。融合对于计算机、电视和电话网络而言，都发挥着日益重要的作用，并具有日益深远的影响。本书一开始就指出这种发展趋势，并从各种通信网络最通用的角度展示计算机网络的主要工作机制，比如多路复用、交换和路由。

本书读者对象

本书的内容已经在各类听众中得到了成功检验。这些听众包括具有不同经验和专业兴趣的学生，其中有大学的本科生和研究生、IT部门的负责人、网络管理员和集成制造商等。对初学者，本书可以为进一步的学习打下坚实的基础；而对专业人士，本书可以帮助其组织和完善相关知识。

本书主要面向那些希望能够系统地掌握网络理论和实践知识的本科生和研究生。

我们希望本书也能对某些专业人士提供帮助。他们对IT技术的了解仅始于在用连接到因特网的PC机进行实际操作时所掌握的网络运行的基本常识。那些想掌握基础知识的读者能够通过自学本书对网络运行做进一步的理论研究。

本书对有经验的网络专业人员也是有用的。他们可以在本书中接触到过去在实际工作过程中未曾遇到的新技术，也可借助书中的内容对自身已有的知识进行整理。如果作为实际工作的参考书，本书也很合适，因为在书中可以找到特定协议、帧格式等的详细描述。此外，本书提供了准备Cisco认证考试（如CCNA、CCNP、CCDP和CCIP）所需要的理论知识。

全书的结构

本书由五部分，总计24章组成。

- 第一部分，网络基础。这部分涵盖了学习计算机网络螺旋过程的“第一圈”。我们知道，认知过程总是螺旋式上升的。立即就对某个复杂的现象全面理解是不可能的。相反，任何复杂的现象必须从各个不同的角度（总体和细节）进行学习，可能随时需要重新温习某些似乎已经理解的材料，而每一次螺旋认识过程的轮回都会让人收获新的信息。第一部分由7章组成，描述了主要的、也是最重要的原理和体系结构解决方案。它们是本书后面可能遇到的所有现代网络技术的基础。从网络融合的角度出发，我们以最一般化的观点介绍计算机网络的交换、多路复用、路由、寻址和体系结构等原理，并与电话、通信载波、无线电以及电视网络等其他电信网络的类似原理相比较。本部分以涉及分组交换网络中服务质量（QoS）问题的一章结束。因此，长期以来被看作网络技术一个特殊分支的QoS成为了构建计算机网络的基本原理之一。
- 第二部分，物理层技术。本部分有四章：第8章，传输链路；第9章，数据编码和多路复用；第10章，无线传输；第11章，传输网络。其中，头两章描述了各种类型的传输链路，并对网络中发送各种离散信息的现代方法提供了详细的信息。书中这部分材料使读者不用花太多的时间去阅读大量的专业文献就能掌握最精简的必需知识。这些知识范围包括信息理论、光谱分析、物理和逻辑数据编码以及差错检测和校正等。第10章重点讲解目前日益流行的无线数据传输技术。高噪声和复杂的波传播路径要求在无线通信链路上采用特殊的信号编码和传输方法。第11章包括如准同步数字系列（PDH）、SDH/SONET和密集波分多路复用等技术，这些构成了全球电信网络物理链路的基础架构。计算机网络和电话网正是在传输网络形成的信道之上运行的。
- 第三部分，局域网。本部分详细介绍了实际中主要的局域网（LAN）技术，包括以太网、令牌环、光纤分布式数据接口（FDDI）以及新兴的高速技术。当前的LAN中，一种技术或者更确切地说是一组技术——以太网——占有统治地位。当然，我们也相比其他网络更详细地讨论了以太网技术。第12章涵盖了经典的10 Mb/s以太网技术，第13章描述基于共享媒体的高速以太网，即快速以太网和千兆以太网。第14章描述其他一些使用共享媒体的LAN技术，包括令牌环、FDDI和两种无线网络技术——IEEE 802.11局域网和蓝牙个人网络。该部分的最后两章，第15章和第16章，主题是交换LAN。前一章介绍交换LAN的主要原理：LAN交换机运行的算法、全双工的LAN协议以及LAN交换机实现的特点。第16章研究这类网络的扩展能力，包括基于生成树算法的备份链路和虚拟LAN技术。
- 按照开放式系统互连参考模型指出的逻辑关系，在专注于物理层和数据链路层的部分后的第四部分，我们主要讲解了网络层技术，它使大量不同的本地网络组合成为一个统一的互联网络成为可能。由于IP是网络层协议中无可争议的领先者，因此我们在本书中对其重点关注。第17章描述IP寻址的各个方面：本地地址、网络地址及符号地址的映射方法；使用网络掩码的方法聚合IP地址的方法以及IP节点自动配置的方法。第18章详细分析了与分组转发和分片相关的IP操作，描述了路由表的通用格式，并提供了在不同种类软硬件路由器中特定实现的例子。当描述新版本IP——IPv6——的具体特性时，我们包括了详细的新寻址方法及引入IP报头格式的主要变化。第19章以传输控制协议（TCP）和用户数据报协议（UDP）的学习开始，这两种协议在应用和网络运输基础设施间扮演了中间角色。接着，包括了路由信息协议（RIP）、开放最短路径优先（OSPF）协议和边界网关协议（BGP）。文中提供的资料分析了这些协议的应用范围以及它们组合使用的可能性。本章以因特网控制报文协议的介绍结束，

它是通知发送方为什么它的分组没有被传递给目的节点的方法。第20章描述了路由器的种类和主要特性、它们内部组织的变化以及在同一设备上组合交换和路由功能的方法——第三层交换机。第四部分对TCP/IP协议栈的全面介绍对了解IP网络是十分有价值的。

- 第五部分，广域网。本部分由四章组成。本书前一部分介绍的IP技术使构建不同类型的互联网络（包括局域网和全球网）成为可能。此外，还有一些网络技术是以专为广域网开发的虚电路技术为基础。在第21章中包括了这些在帧中继和ATM网络中实现的技术。虚电路代表了一种可替代作为以太网和IP网络基础的分组转发数据报方法的技术。这两类主要数据传输技术间的竞争已存在多时，实际上从分组交换网络发展以来就已经开始。第22章考虑了利用IP技术建立广域网的各个方面。多协议标记交换（MPLS）是一种在综合IP和虚电路技术领域的创新。它处于IP层和如ATM、帧中继或以太网这些技术的层之间，将它们联合到统一有效的运输系统中。本章以对基于简单网络管理协议的网络管理系统介绍结束。简单网络管理协议不但广泛应用于IP路由器控制，而且也应用于各种类型电信设备的控制。第23章研究了为网络用户提供主干网高速接入的各种方法。最有效的技术是利用现有的电缆基础设施（比如，电话网络本地回路上的不对称用户数字线）或使用有线电视系统的电缆调制解调器。另一种替代的解决方法是无线接入，它可以是移动的，也可以是固定的。第24章作为本部分也是全书的结束。它专注于网络运输系统的安全问题。该章包括各种类型的虚拟专用网（VPN），特别是基于安全版本IP（IPSec）的虚拟专用网和当前最为流行的一种VPN技术，即MPLS VPN。

我们力图让读者尽可能有效地利用本书。每一章都有一个小结，总结了该章的主要思想、论题和结果。这将有助于读者不再因为大量有用的事实和细节而忽略主要原理。在每一章的结尾还有复习题和练习题，旨在检验读者阅读本章而获得的知识。在某些情况下，这些练习题有特殊的意义，因为它能够让读者更好地理解某些思想。

支持的Web站点

感兴趣的读者能够在我们的网站http://www.olifer.co.uk上找到更多的信息。我们希望这个站点能够成为学生、教师以及网络专业人士学习本书的补充内容。当然，网站内容会不断更新。一开始，我们计划提供下述材料：

- 收录了本书所有的插图。
- 本书每一章的复习题和练习题答案。
- 本书的导读，旨在帮助培训者根据本书来创立如下课程：无线网络、IP介绍、服务质量以及远程访问等。这些导读简要地描述了包含适当材料的各章节的排列顺序，有的还提供有关教学方法的某些提示。
- 可以用来作为学期论文选题的案例研究。
- 本书中所涉及主题的诸多因特网链接。
- 读者意见、观点、问题以及对印刷错误或其他错误的指正。

目　录

译者序
译者简介
前言

第一部分　网络基础

第1章　计算机网络的发展 ……2
1.1　引言 ……2
1.2　计算机网络的起源 ……2
1.2.1　计算机网络是计算和通信技术发展的产物 ……2
1.2.2　批处理系统 ……3
1.2.3　多终端系统：计算机网络的原型 …3
1.3　第一代计算机网络 ……4
1.3.1　第一代广域网（WAN）……4
1.3.2　第一代局域网（LAN）……6
1.4　网络融合 ……7
1.4.1　LAN和WAN的融合 ……7
1.4.2　计算机网络和电信网络的融合 ……9
小结 ……10
复习题 ……10
练习题 ……11
第2章　网络设计的一般原理……12
2.1　引言……12
2.2　共享计算机资源的问题……12
2.2.1　计算机与外部设备间的交互作用 …12
2.2.2　两个计算机间最简单的交互作用 …14
2.2.3　网络应用程序 ……16
2.3　使用通信链路的物理数据传输的问题……17
2.3.1　编码 ……18
2.3.2　物理链路的特性 ……19
2.4　多台计算机交互的问题……20
2.4.1　物理链路的拓扑 ……20
2.4.2　网络节点的编址 ……22
2.4.3　交换 ……24
2.5　通用的交换问题……24
2.5.1　流定义 ……24
2.5.2　路由 ……25
2.5.3　数据转发 ……27
2.5.4　多路复用和解多路复用 ……28
2.5.5　共享介质 ……29
2.5.6　交换类型 ……30
小结 ……31
复习题 ……31
练习题 ……32
第3章　分组和电路交换 ……33
3.1　引言……33
3.2　电路交换……33
3.2.1　连接建立 ……34
3.2.2　建立请求的阻塞 ……34
3.2.3　保证带宽 ……34
3.2.4　多路复用 ……35
3.2.5　传送突发流量的低效率 ……39
3.3　分组交换……37
3.3.1　缓存与队列 ……37
3.3.2　分组转发方法 ……39
3.3.3　数据报传输 ……39
3.3.4　逻辑连接 ……41
3.3.5　虚电路 ……41
3.3.6　电路交换网络与分组交换网络的比较 ……43
3.4　在共享介质网络中的分组交换……48
3.4.1　介质共享的原理 ……48
3.4.2　LAN结构的理由 ……49
3.4.3　LAN的物理构造 ……50
3.4.4　共享介质网络的逻辑构造 ……51
3.4.5　作为标准技术例子的以太网 ……53
小结 ……54
复习题 ……55
练习题 ……55

第4章 网络体系结构与标准……57
4.1 引言……57
4.2 网络节点互动的分解……57
4.2.1 多层的方法……57
4.2.2 协议和协议栈……59
4.3 OSI模型……60
4.3.1 OSI模型的一般特性……60
4.3.2 物理层……62
4.3.3 数据链路层……62
4.3.4 网络层……63
4.3.5 运输层……66
4.3.6 会话层……67
4.3.7 表示层……67
4.3.8 应用层……67
4.3.9 OSI模型和电路交换网络……67
4.4 网络标准……67
4.4.1 开放系统的概念……68
4.4.2 标准的类型……68
4.4.3 因特网标准……69
4.4.4 通信协议的标准栈……69
4.4.5 流行协议栈与OSI模型间的对应关系……74
4.5 信息和运输服务……75
4.5.1 网络元素的协议分布……76
4.5.2 运输系统的辅助协议……77
小结……78
复习题……78
练习题……79
第5章 网络的例子……80
5.1 引言……80
5.2 电信网络的一般结构……80
5.2.1 接入网……81
5.2.2 主干……81
5.2.3 数据中心……81
5.3 电信运营商网络……82
5.3.1 服务……82
5.3.2 客户……83
5.3.3 基础结构……84
5.3.4 覆盖的范围……85
5.3.5 不同类型运营商间的关系……85
5.4 公司网络……86
5.4.1 部门网……87
5.4.2 楼宇或校园网……87
5.4.3 企业范围的网络……88
5.5 因特网……90
5.5.1 因特网的独特性……90
5.5.2 因特网的结构……91
5.5.3 因特网的边界……93
小结……95
复习题……95
练习题……96
第6章 网络的特性……97
6.1 引言……97
6.2 特性的类型……97
6.2.1 主观质量特性……97
6.2.2 网络特性和要求……98
6.2.3 时间尺度……98
6.2.4 服务水平约定……99
6.3 性能……99
6.3.1 理想的网络……99
6.3.2 分组延迟的特性……101
6.3.3 信息率的特性……103
6.4 可靠性……104
6.4.1 分组丢失特性……104
6.4.2 可用性和容错……104
6.4.3 可替换的路由……105
6.4.4 数据重传和滑动窗口……106
6.5 安全……108
6.5.1 计算机和网络安全……109
6.5.2 数据保密性、完整性和可用性……109
6.5.3 网络安全服务……110
6.6 仅用于服务提供商的特性……111
6.6.1 可扩展性和可延拓性……111
6.6.2 可管理性……112
6.6.3 兼容性……112
小结……112
复习题……113
练习题……113
第7章 保证服务质量的方法……114
7.1 引言……114
7.2 应用与QoS……114
7.2.1 不同类型应用的QoS要求……114
7.2.2 信息率的可预测性……115
7.2.3 应用对分组延迟的敏感性……116

7.2.4 应用对分组丢失的敏感性 ……116
7.2.5 应用类别 ……117
7.3 队列分析 ……117
7.3.1 M/M/1模型 ……118
7.3.2 作为分组处理模型的M/M/1 ……119
7.4 QoS机制 ……121
7.4.1 在低负载方式下运行 ……121
7.4.2 不同的服务类别 ……121
7.5 队列管理算法 ……122
7.5.1 FIFO算法 ……122
7.5.2 优先权排队 ……122
7.5.3 加权排队 ……124
7.5.4 混合的排队算法 ……125
7.6 反馈 ……125
7.6.1 目的 ……125
7.6.2 反馈参与者 ……126
7.6.3 反馈信息 ……127
7.7 资源预留 ……128
7.7.1 资源预留和分组交换 ……128
7.7.2 基于预留的QoS系统 ……131
7.8 流量工程 ……133
7.8.1 传统路由方法的不足 ……133
7.8.2 流量工程的思想 ……134
7.8.3 不同流量类别的流量工程 ……136
小结 ……137
复习题 ……137
练习题 ……137

第二部分 物理层技术

第8章 传输链路 ……140
8.1 引言 ……140
8.2 分类 ……140
8.2.1 传输网、电路和链路 ……140
8.2.2 介质 ……141
8.2.3 传输设备 ……142
8.3 传输链路特性 ……143
8.3.1 通信链路中信号的频谱分析 ……143
8.3.2 衰减与阻抗 ……145
8.3.3 抗噪声与传输可靠性 ……146
8.3.4 带宽与容量 ……148
8.3.5 比特与波特 ……149
8.3.6 带宽与容量间的相关性 ……150
8.4 电缆类型 ……151
8.4.1 非屏蔽和屏蔽双绞线 ……151
8.4.2 同轴电缆 ……152
8.4.3 光缆 ……153
8.4.4 楼宇的结构化布线系统 ……154
小结 ……155
复习题 ……155
练习题 ……156
第9章 数据编码和多路复用 ……157
9.1 引言 ……157
9.2 调制 ……157
9.2.1 传输模拟信号时的调制 ……157
9.2.2 传输离散信号时的调制 ……157
9.2.3 组合调制方式 ……158
9.3 数字化模拟信号 ……160
9.3.1 脉冲编码调制 ……160
9.3.2 数字化声音 ……161
9.4 编码方法 ……161
9.4.1 选择编码方法 ……161
9.4.2 电平不归零码 ……162
9.4.3 双极标记交替反转编码 ……163
9.4.4 “1”翻转的不归零码 ……163
9.4.5 双极脉冲编码 ……164
9.4.6 曼彻斯特编码 ……164
9.4.7 2B1Q电平码 ……164
9.4.8 冗余码 ……165
9.4.9 扰频 ……165
9.4.10 数据压缩 ……167
9.5 差错检测与校正 ……168
9.5.1 差错检测技术 ……168
9.5.2 差错校正 ……169
9.6 多路复用和交换 ……169
9.6.1 基于FDM和WDM的电路交换 ……170
9.6.2 基于TDM的电路交换 ……171
9.6.3 信道运行的双工方式 ……172
小结 ……173
复习题 ……173
练习题 ……174
第10章 无线传输 ……175
10.1 引言 ……175
10.2 无线介质 ……175

10.2.1 无线通信的优点 ……175
10.2.2 无线链路 ……176
10.2.3 电磁频谱 ……176
10.2.4 电磁波的传播 ……177
10.2.5 许可 ……178
10.3 无线系统……179
10.3.1 点对点系统 ……179
10.3.2 点对多点系统 ……180
10.3.3 多点对多点系统 ……181
10.3.4 卫星系统 ……181
10.3.5 同步卫星 ……183
10.3.6 媒体和中低轨道卫星 ……183
10.4 扩频技术……184
10.4.1 跳频扩频 ……185
10.4.2 直接序列扩频 ……186
10.4.3 码分多路访问 ……187
小结 ……188
复习题 ……189
练习题 ……189
第11章 传输网络 ……190
11.1 引言……190
11.2 PDH网 ……190
11.2.1 速率层次 ……190
11.2.2 多路复用方法 ……191
11.2.3 PDH技术的局限性 ……192
11.3 SONET/SDH网 ……193
11.3.1 速率层次与多路复用方法 ……193
11.3.2 设备类型 ……195
11.3.3 协议栈 ……196
11.3.4 STM-N帧 ……196
11.3.5 典型的拓扑 ……198
11.3.6 保证网络抗毁性的方法 ……199
11.4 DWDM网络 ……202
11.4.1 运行原理 ……203
11.4.2 光纤放大器 ……204
11.4.3 典型的拓扑 ……205
11.4.4 光添加/丢弃多路复用器 ……206
11.4.5 光交叉连接器 ……207
11.5 案例学习……208
小结 ……209
复习题 ……210
练习题 ……211

第三部分 局 域 网

第12章 以太网 ……214
12.1 引言……214
12.2 LAN协议的一般特性 ……214
12.2.1 标准拓扑和共享介质 ……214
12.2.2 LAN协议栈 ……215
12.2.3 IEEE 802.x标准的结构 ……219
12.3 CSMA/CD ……220
12.3.1 MAC地址 ……220
12.3.2 介质访问和数据传输 ……221
12.3.3 冲突 ……222
12.3.4 路径延迟值和冲突检测 ……223
12.4 以太网帧格式……224
12.4.1 802.3/LLC ……225
12.4.2 原始802.3/Novell 802.3帧 ……226
12.4.3 以太网DIX/以太网II帧 ……226
12.4.4 以太网SNAP帧 ……226
12.4.5 使用各种类型的以太网帧 ……226
12.5 以太网的最好性能……227
12.6 以太网物理介质规范……228
12.6.1 10Base-5 ……229
12.6.2 10Base-2 ……230
12.6.3 10Base-T ……231
12.6.4 光纤以太网 ……233
12.6.5 冲突域 ……234
12.6.6 10Mb/s以太网标准的公共特性 …234
12.7 案例学习……234
小结 ……237
复习题 ……238
练习题 ……239
第13章 高速以太网 ……241
13.1 引言……241
13.2 快速以太网……241
13.2.1 历史概述 ……241
13.2.2 快速以太网的物理层 ……242
13.2.3 100Base-FX/TX/T4规范 ……243
13.2.4 使用中继器构建快速以太网段的规则 ……245
13.2.5 100VG-AnyLAN的特殊性质 ……246
13.3 千兆以太网……247
13.3.1 历史概述 ……247

13.3.2 问题 ……248
13.3.3 保证200m直径的网络 ……249
13.3.4 802.3z物理介质规范 ……249
13.3.5 基于5类双绞线的千兆以太网 …249
小结 ……250
复习题 ……251
练习题 ……251
第14章 共享介质的LAN ……253
14.1 引言 ……253
14.2 令牌环 ……253
14.2.1 令牌传递访问 ……253
14.2.2 令牌环物理层 ……255
14.3 FDDI ……256
14.3.1 主要的FDDI特性 ……256
14.3.2 FDDI容错 ……257
14.4 无线LAN ……259
14.4.1 无线LAN的特殊性质 ……259
14.4.2 IEEE 802.11协议栈 ……261
14.4.3 802.11 LAN的拓扑 ……262
14.4.4 访问共享介质 ……263
14.4.5 安全 ……265
14.5 PAN与蓝牙 ……266
14.5.1 PAN的特殊性质 ……266
14.5.2 蓝牙的体系结构 ……266
14.5.3 蓝牙协议栈 ……268
14.5.4 蓝牙帧 ……269
14.5.5 蓝牙如何运作 ……270
14.6 共享介质LAN的设备 ……270
14.6.1 网络适配器的主要功能 ……271
14.6.2 集中器的主要功能 ……272
14.6.3 自动分隔 ……273
14.6.4 反相链路的支持 ……274
14.6.5 保护以防未授权访问 ……274
14.6.6 多段集中器 ……275
14.6.7 集中器设计 ……276
小结 ……277
复习题 ……278
练习题 ……279
第15章 交换LAN基础 ……280
15.1 引言 ……280
15.2 使用网桥和交换机的逻辑网络结构…280
15.2.1 共享介质LAN的优点与不足 ……280
15.2.2 逻辑网络结构的优点 ……281
15.2.3 IEEE 802.1D标准的透明网桥算法 ……283
15.2.4 交换机LAN的拓扑局限性 ……287
15.3 交换机 ……288
15.3.1 交换机的特殊性质 ……288
15.3.2 无阻塞的交换机 ……292
15.3.3 克服拥塞 ……292
15.3.4 数据链路层协议的翻译 ……293
15.3.5 流量过滤 ……294
15.3.6 交换机体系结构和设计 ……294
15.3.7 交换机的性能特性 ……297
15.4 全双工LAN协议 ……299
15.4.1 在全双工模式运行中引入MAC层的变化 ……299
15.4.2 在全双工模式中拥塞控制的问题 ……300
15.4.3 10G以太网 ……301
小结 ……303
复习题 ……303
练习题 ……304
第16章 交换LAN的高级特性 ……305
16.1 引言 ……305
16.2 生成树算法 ……305
16.2.1 必要的定义 ……306
16.2.2 构建生成树的三步过程 ……307
16.2.3 STA的优点和不足 ……309
16.3 LAN中的链路聚合 ……309
16.3.1 干线与逻辑信道 ……309
16.3.2 消除帧的生育 ……311
16.3.3 端口选择 ……312
16.4 虚拟LAN ……314
16.4.1 VLAN目的 ……315
16.4.2 构建基于一个交换机的VLAN …316
16.4.3 构建基于多个交换机的VLAN …316
16.5 LAN中的服务质量 ……319
16.6 网桥和交换机的局限性 ……321
16.7 案例学习 ……321
小结 ……322
复习题 ……323

第四部分 TCP/IP网际互联

第17章 TCP/IP网络中的寻址 ……326
17.1 引言……326
17.2 TCP/IP栈的地址类型 ……326
17.2.1 本地地址 ……326
17.2.2 IP网络地址 ……327
17.2.3 域名 ……327
17.3 IP地址格式……328
17.3.1 IP地址的分类 ……328
17.3.2 特殊的IP地址 ……329
17.3.3 在IP地址中使用掩码 ……330
17.4 IP地址分配顺序……331
17.4.1 自治网络中的地址分配 ……331
17.4.2 集中式的地址分配 ……332
17.4.3 寻址和CIDR……332
17.5 将IP地址映射到本地地址……333
17.5.1 ARP ……334
17.5.2 代理ARP ……337
17.6 DNS ……338
17.6.1 平面符号名称 ……338
17.6.2 层次式符号名称 ……338
17.6.3 DNS的操作方式 ……339
17.6.4 反向搜索区域 ……341
17.7 DHCP ……341
17.7.1 DHCP方式 ……342
17.7.2 动态地址分配算法 ……343
小结 ……344
复习题 ……345
练习题 ……346
第18章 因特网协议 ……347
18.1 引言……347
18.2 IP分组格式……347
18.3 IP路由方法……349
18.3.1 简化的路由表结构 ……350
18.3.2 端节点上的路由表 ……352
18.3.3 搜索不含掩码的路由表 ……352
18.3.4 不同格式路由表的例子 ……353
18.3.5 在路由表中记录的来源和类型 …356
18.3.6 不带掩码的IP路由的例子 ……357
18.4 使用掩码的路由……360
18.4.1 构造一个带同样长度掩码的网络 ……360
18.4.2 考虑掩码的表查找算法 ……362
18.4.3 使用可变长的掩码 ……363
18.4.4 复用地址空间 ……365
18.4.5 路由和CIDR……368
18.5 IP分组的分片……369
18.5.1 MTU作为一个技术参数 ……370
18.5.2 分片参数 ……370
18.5.3 分片和组装分组的过程 ……371
18.5.4 分片的例子 ……371
18.6 IPv6……372
18.6.1 TCP/IP栈的新方向 ……372
18.6.2 可延拓的寻址系统 ……373
18.6.3 灵活的头格式 ……377
18.6.4 减少路由器的负荷 ……378
小结 ……378
复习题 ……379
练习题 ……380
第19章 TCP/IP栈的核心协议 ……381
19.1 引言……381
19.2 TCP和UDP运输层协议……381
19.2.1 端口 ……381
19.2.2 UDP ……382
19.2.3 TCP段格式 ……384
19.2.4 作为TCP可靠性基础的逻辑连接…385
19.2.5 序列号和确认号 ……387
19.2.6 接收端窗口 ……388
19.2.7 累积确认原则 ……389
19.2.8 确认超时 ……390
19.2.9 控制接收端窗口 ……390
19.3 路由协议……391
19.3.1 路由协议的分类 ……391
19.3.2 路由信息协议 ……395
19.3.3 开放最短路径优先 ……400
19.3.4 边界网关协议 ……402
19.4 因特网控制报文协议……404
19.4.1 ICMP报文的类型 ……405
19.4.2 回送请求/响应报文的格式：Ping实用程序 ……406
19.4.3 错误报文格式：Traceroute实用程序 ……406
小结 ……408
复习题 ……409

练习题 ……410
第20章 IP路由器的高级特性……411
20.1 引言……411
20.2 过滤……411
20.2.1 用户流量过滤 ……411
20.2.2 路由公告过滤 ……413
20.3 IP QoS……414
20.3.1 IntServ和Difdserv QoS模型 ……414
20.3.2 令牌桶算法 ……415
20.3.3 随机早期检测 ……416
20.3.4 集成服务框架和RSVP ……417
20.3.5 区分服务框架 ……419
20.4 网络地址转换……422
20.4.1 地址转换的原因 ……422
20.4.2 传统的NAT ……422
20.4.3 基本的NAT ……423
20.4.4 地址和端口转换 ……424
20.5 路由器……426
20.5.1 路由器功能 ……426
20.5.2 路由器按应用范围的分类 ……427
小结 ……431
复习题 ……432
练习题 ……432

第五部分 广 域 网

第21章 虚电路WAN ……436
21.1 引言……436
21.2 虚电路技术……436
21.2.1 交换虚电路 ……436
21.2.2 永久虚电路 ……439
21.2.3 与数据报技术的比较 ……439
21.3 X.25网络 ……440
21.3.1 X.25网络的结构和目的 ……440
21.3.2 X.25网络寻址 ……441
21.3.3 X.25网络协议栈 ……441
21.4 帧中继网……442
21.4.1 帧中继协议栈 ……443
21.4.2 QoS支持 ……445
21.5 ATM技术 ……447
21.5.1 ATM运行的主要原理 ……447
21.5.2 ATM协议栈 ……450
21.5.3 ATM适配层 ……451
21.5.4 ATM协议 ……452
21.5.5 ATM协议服务和流量控制的种类 ……454
小结 ……457
复习题 ……458
练习题 ……458
第22章 IP WAN……460
22.1 引言……460
22.2 纯IP WAN ……460
22.2.1 IP WAN结构 ……460
22.2.2 HDLC族的协议 ……462
22.2.3 点到点协议 ……464
22.2.4 IP路由器使用的租用线 ……465
22.3 在ATM或帧中继上的IP ……466
22.3.1 IP和ATM层间的通信 ……466
22.3.2 配置路由器接口 ……467
22.4 多协议标记交换……468
22.4.1 在同一设备中组合交换和路由 ……468
22.4.2 LSR和数据转发表 ……468
22.4.3 标记交换路径 ……470
22.4.4 MPLS头和数据链路技术 ……471
22.4.5 标记栈 ……472
22.4.6 MPLS应用领域 ……474
22.4.7 MPLS内部网关协议 ……474
22.4.8 MPLS流量工程 ……476
22.5 网络管理……479
22.5.1 网络管理系统的目的 ……479
22.5.2 网络管理问题的功能组 ……479
22.5.3 网络管理系统的体系结构 ……480
22.5.4 基于SNMP的管理系统标准 ……483
22.5.5 SNMP MIB结构 ……483
22.5.6 SNMP报文格式 ……486
22.5.7 RMON MIB规范 ……487
小结 ……489
复习题 ……489
练习题 ……490
第23章 远程访问 ……491
23.1 引言……491
23.2 远程访问的方法……491
23.2.1 客户和终端设备的类型 ……492
23.2.2 在本地环路的信息多路复用 ……493
23.2.3 远程节点方式 ……495
23.2.4 远程控制方式Telnet ……496

23.3 拨号模拟访问……497
23.3.1 电话网运行的原理 ……497
23.3.2 通过电话网远程访问 ……499
23.3.3 调制解调器 ……500
23.4 用ISDN拨号访问 ……502
23.4.1 ISDN的目的和结构……502
23.4.2 BRI和PRI接口 ……503
23.4.3 ISDN协议栈……504
23.4.4 用ISDN进行数据传输……506
23.5 XDSL技术 ……508
23.6 用有线电视访问……510
23.7 无线访问……511
小结 ……512
复习题 ……512
练习题 ……513
第24章 安全的运输服务……514
24.1 引言……514
24.2 IPSec受保护的信道服务 ……514
24.2.1 受保护的信道的服务层次 ……514
24.2.2 IPSec协议间的功能分配……515
24.2.3 IPSec中的加密……516
24.2.4 安全关联 ……517
24.2.5 运输和隧道模式 ……518
24.2.6 AH协议……520
24.2.7 ESP协议 ……521
24.2.8 安全数据库 ……521
24.3 虚拟专用网服务……523
24.3.1 VPN定义 ……523
24.3.2 VPN评价和比较的准则 ……524
24.3.3 在流量分离基础上的VPN ……525
24.3.4 IPSec VPN ……527
24.4 MPLS VPN……527
24.4.1 完全连接和绝对隔离 ……528
24.4.2 MPLS VPN部件 ……529
24.4.3 路由信息的分离 ……530
24.4.4 用MP-BGP连接站点 ……531
24.4.5 地址空间的无关性 ……532
24.4.6 MP-BGP路由广告的生成 ……534
24.4.7 在MPLS VPN上的分组转发 ……534
24.4.8 形成VPN拓扑的机制 ……535
24.4.9 安全水平 ……537
小结 ……537
复习题 ……537
练习题 ……538
结束语 展望未来 ……539
参考文献与推荐阅读的书 ……541

第一部分 网络基础

认知的过程具有螺旋式上升的性质，我们不可能立刻理解和意识到复杂的现象，为了更好地认识这些现象，我们必须从不同的角度加以考虑，既需要从整体和部分的角度，又需要各自独立地并与其他现象联系起来，逐渐地增加我们的知识。此外，我们还需要不时地回到那些似乎已经理解了的概念。在每个螺旋的转折处，我们将对这一现象的本质有更好的理解。对此，一个好的方法是首先学习一个特定知识领域的通用原理，然后仔细地研究这些原理如何在特定的方法、技术、结构中实现。

本书的开篇就是学习计算机网络的第一个螺旋。它介绍了构成所有当代网络技术基础的主要原理和体系结构，这些内容在本书的后续章节还会进一步介绍。根据网络融合的概念，我们将从最基本和最普遍的角度，介绍交换、多路复用、路由、寻址、网络体系结构的原理。我们通过将计算机网络的原理与其他通信网络（例如，电话网络、传输网络、广播和电视网络）的类似原理相比较来进行介绍。

本部分的最后一章介绍了分组交换网络中的服务质量（QoS）问题。作为下一代公共网络发展的基础，计算机网络能提供各种信息服务，传输数据、语音和视频流量。计算机网络的这一新角色造成了服务质量衡量方法在几乎所有通信技术中的应用。因此，QoS虽然在很长的时间内被认为属于专门的、高级的网络技术领域，但是，现在它已经成为了建造计算机网络所使用的基础概念。

在详细学习完特定技术后，回到本书的第一部分将会非常有用（同时，作者希望也会非常有趣）。这一学习过程的迭代将使读者更好地理解计算机网络的基本操作原理，以及这些原理在不同技术中的实现。

第1章 计算机网络的发展

1.1 引言

学习任何科学或技术领域的发展不仅会增强对自然科学的好奇心，而且还会使你对这一领域的主要成就有更深一步的认识，帮助你了解现有的发展趋势，评估最新的发展。相对而言，计算机网络出现得较晚，大约出现在20世纪60年代末。他们从更古老的、更广泛使用的电话网络中继承了很多有用的特性。这并不令人惊奇，因为计算机和电话都是普遍使用的通信工具。

然而，计算机网络确实给通信领域带来了一些新的东西——数千年的人类文明积累下来的无穷无尽的信息。这些存储的信息越来越快地积累起来。这一现象在90年代中期变得更加明显——因特网（Internet）的快速发展显示出人们非常重视免费的、匿名的信息访问，以及即时的、书面的通信。

计算机网络对其他电信网络的影响造成了网络融合，这一过程在Internet产生之前就已经开始，这种融合的第一个标志就是电话网络上的数字语音传输。网络融合的最新迹象出现在计算机网络最新开发的服务上。例如，网络电话（VoIP）、无线电广播和电视服务，以前，它们更多地表现在电话网络、广播网络、电视网络。这种融合的过程还在继续，虽然它的未来何去何从尚不清楚，但如果能了解本章所介绍的计算机网络的发展，便可以更好地理解计算机网络研究者所面临的主要问题。

1.2 计算机网络的起源

1.2.1 计算机网络是计算和通信技术发展的产物

本章所介绍的**计算机网络**（computer network）显然不是人类文明所建立的唯一的网络。最早覆盖很大领域并服务许多终端的网络可能是古罗马时代的供水系统。虽然不同的网络会有很大的不同，但它们都有一些共同点。例如，在电力网络和大规模计算机网络之间总是可以进行一些类比——计算机网络上的信息源对应于发电厂；计算机网络的通信链路对应于高压电线；接入网与变电站相似。同时，在计算机网络和电力网络中，我们都可以找到用户终端——计算机网络中的用户工作站和电力网络中的家用电器。

计算机网络，有时也被称为**数据传输网络**（datacom or data-transmission network），是现代文明的两种最重要的科学和技术分支——计算技术和通信技术发展的结果。

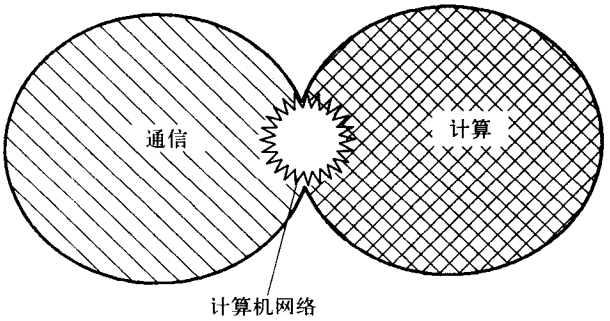


图1-1 计算机网络的发展——计算技术和通信技术的结合

一方面（图1-1），计算机网络代表了分布式计算系统的一个特例。在这一特例下，一组互相协调的计算机通过自动交换数据来完成一些互相交织的任务。计算机网络也可以被认为是一种长距离传送信息的方式。为了实现这一目标，计算机网络实现了在许多通信系统中广泛使用的数据编码和多路复用。

1.2.2 批处理系统

首先，我们介绍计算机网络的起源。20世纪50年代的计算机庞大而又昂贵，只供少量特殊用户使用。通常，这些巨大的机器会占据整座建筑物。这些计算机不能与用户交互，它们批处理大量的工作，再把结果返还回来。

批处理系统（batch-processing system）通常基于大型机，它们往往是一些处理能力强大、可靠性高的通用计算机。用户要准备打孔卡片，这些卡片记载着数据和程序。随后，用户将这些卡片送到计算中心。操作员将这些卡片输入到计算机中，用户在一天后得到打印输出的结果（图1-2）。因此，一张带有错误的打孔卡片往往会耽误至少24小时。

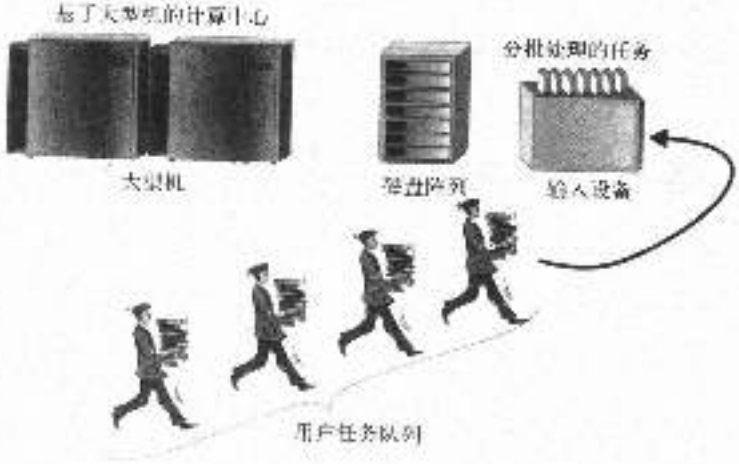


图1-2 基于大型机的集中系统

显然，从终端用户的角度来看，可交互的操作方式可以使他们能够在终端上动态地管理他们所需处理的数据，并且大为方便。在计算机系统发展的早期，用户的利益被大量地忽略。计算机系统最昂贵的部件——处理器，其性能被认为是重中之重，为此，有时需要付出牺牲用户利益的代价。

1.2.3 多终端系统：计算机网络的原型

当处理器在1960年初变得更加便宜时，新的组织计算机处理的方法出现了。这些方法使得方便终端用户成为可能。于是，多终端系统出现了（表1-3）。在这种分时系统中，一台计算机由多个用户使用，每个用户操作他们自己的终端，从这个终端上，他们可以和计算机通信。计算机系统对每个用户的响应时间都很短。这样，感觉上像是计算机同时并行地为多个用户提供服务。

终端被移出了计算中心，并被摆上了用户的桌面。尽管处理能力仍然完全是集中式的，但有些功能成了分布式的，例如数据的输入和输出。这种集中式的多终端系统看上去和局域网（LAN）类似。终端用户觉得在终端上工作，就像今天在连接着网络的PC上工作一样，用户可以访问共享文件和其他的外部设备。由于用户可以在任何时候启动一个需要的程序，并且立刻得到结果，所以，从感觉上，用户觉得像是单独地使用这台电脑。（有些用户甚至相信所有的计算都是在计算机显示器内完成的。）

通过分时提供工作的多终端系统是局域网发展的第一步。

然而，在局域网真正出现之前，这种技术的发展还有很长的路要走。虽然多终端系统与分布式系统非常相像，但多终端系统本质上还是集中式的数据处理。

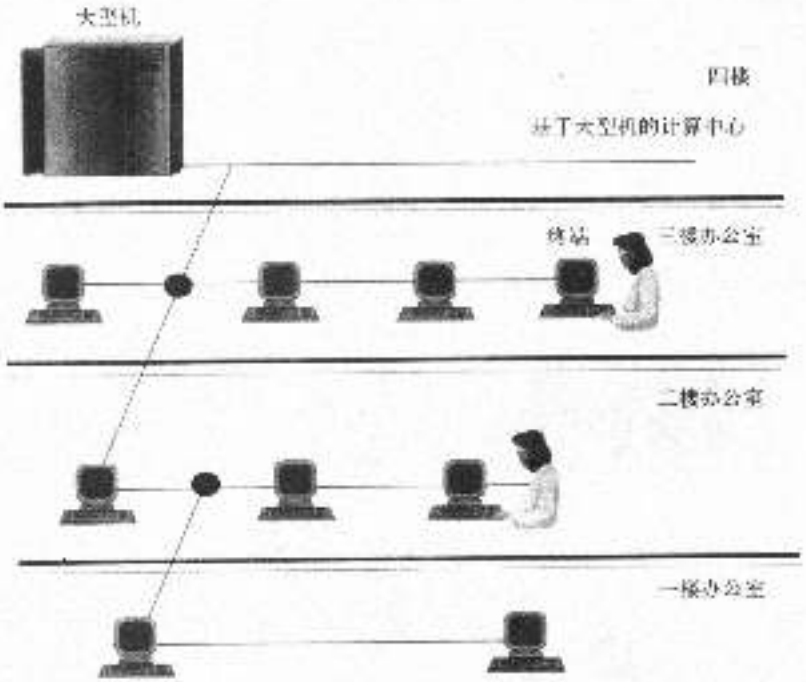


图1-3　多终端系统——计算机网络的原型

很多组织机构对局域网并没有迫切的需要。在一幢建筑物内，没有什么东西可以用网络来连接。许多公司甚至买不起多于一台的计算机。在这一时期，所谓的格罗斯定理（根据赫伯特·格罗斯来命名）是正确的，它经验性地代表了那个时代的技术发展水平。根据这一定理，计算机系统的价格增量是系统计算能力增量的平方项。因此，购买一台高性能机器比购买两台低性能机器更加合算，因为两台低性能机器的总的计算能力远远地低于那台贵的计算机。

1.3　第一代计算机网络

1.3.1　第一代广域网（WAN）

相反地，连接两台远距离计算机的需求已经呈现出来。这些需求从处理一些简单的任务开始，例如，在一个远程终端能够访问另一台坐落在几百英里、甚至几千英里外的计算机。通过电话线，可以使用调制解调器将终端连接到计算机。这些网络允许多个远程用户访问多台超级计算机的共享资源。在分布式系统出现时，不但已经实现了终端和计算机之间的连接，而且还实现了计算机和计算机之间的连接。

计算机变得能够自动地交换数据，这是任何计算机网络的基本工作机制。第一代计算机网络的开发者实现了文件交换、数据库同步、电子邮件以及其他一些现在非常常见的服务。

从时间顺序上说，**广域网**（Wide Area Network, WAN）最先出现。广域网将分布在不同地理位置的计算机连接在一起，有些计算机甚至坐落在不同的城市和国家。

在广域网的发展过程中，人们提出并实现了许多对现代计算机网络起基础作用的想法，例如：

- 通信协议的多层次结构
- 分组交换技术
- 异构网络中的分组路由

虽然广域网从早期广泛使用的长距离网络，例如**电话网络**（telephone network），中继承了许多特性，但它最具创新性的特征是放弃了电路交换的原理。几十年来，电路交换这一方式在电话网络中得到了成功的应用。

在电路交换的整个会话过程中，分配一个固定速率的电路。由于计算机数据具有**突发性流量**（bursty traffic）[⊖]的特点，这种机制显得并不十分有效。（“突发性”指在长时间的等待中夹杂着很小时间段的大量的数据交换）。对于突发性的数据，实验和数学方法都证明了基于分组交换原理的网络比电路交换的网络更有效率。

根据**分组交换**（packet switching）的原理，数据被划分为很小的段，称为**分组**或**包**（packet）。目标主机的地址被嵌入在包头里，这一做法允许每个分组自行在网络上传输。

由于铺设连接远程站点的高质量通信线路非常昂贵，出于各种考虑，第一代的广域网常常使用已有的通信线路。例如，在很长一段时间内，广域网建立在电话网络的基础上。由于使用这样的链路传输速率很低，往往只有几千比特每秒（Kb/s），这种网络所能提供的服务被限制在文件传输（主要是后台模式）和电子邮件。除了传输速率低之外，这些传输信道还有另外一个缺点——传输的信号容易发生错误。因此，使用低质量通信链路的广域网协议不得不具备复杂的数据控制和数据恢复机制。一个典型的例子是20世纪70年代早期开发的X.25网络，那时，连接广域网上计算机和交换机的普遍做法，是从电话公司租借低速率的模拟信道。

1969年，美国国防部提出一项研究，旨在将国防部和研究中心的计算机连接成一个网络。这个称为ARPANET的网络成为了第一代最为广泛使用的广域网，现在它被称做**因特网**（Internet）。

ARPANET将运行不同操作系统、有着不同添加模块的不同类型的计算机连接在一起，它为所有加入其中的计算机实现了共同的通信协议。这些操作系统可以被称为第一代真正的**网络操作系统**（network operation system）。

与多终端系统不同的是，真正的网络操作系统不但允许将系统分布式地提供服务给用户，还分布式地组织数据存储。它们甚至还允许通过链路将计算处理分布到多台不同的电脑上。任何网络操作系统都具备所有本地操作系统所具备的功能，另外，它还会提供一些额外的功能，使系统能够通过网络与其他操作系统进行通信。随着网络技术和计算机硬件的发展，新的网络处理需求逐步出现，实现新的网络功能的计算机软件模块也被渐渐加入到这些操作系统之中。

广域网的发展很大程度上取决于电话网络的发展。

从20世纪60年代晚期开始，数字化的语音传输在电话网络上越来越常见。

这催生了高速数据信道。这些高速数据信道连接自动电话交换机，提供了同时传输几百到几千路通话的能力。研究出了专门的技术用来构造**传输网络**（transmission network）或**主干网**（backbone）。这些网络并不服务于终端用户，相反地，高速的点对点的数据信道以它们作为基础。这些信道连接着另一个服务于终端用户的网络（*覆盖网*）的设备。

刚开始时，传输网仅仅是电话公司内部使用的技术。然而，渐渐地，这些公司将它们部分的数据信道出租给其他公司，这些租借的公司致力于建造他们自己的电话网和广域网。现在，传输网将数据传输速率提高到几百吉比特每秒（Gb/s），在某些情况下甚至可以达到钛比特每秒（Tb/s）。这些网络覆盖了所有主要的工业州。

广域网的多样性和服务质量帮助广域网赶上了局域网。虽然局域网出现得相对较晚，但它却成为了最终的领导者。

⊖ Burst和bursty traffic是被广泛采用的数据通信术语。按照Cisco公司的技术定义，按某种标准或度量burst是可看成是一个单元的信号序列，而bursty traffic指一种不均匀的数据传输模式。

1.3.2　第一代局域网（LAN）

20世纪70年代早期，一件影响计算机网络发展的重要事件发生了。随着计算机部件制造技术发展，大规模集成电路出现了，大规模集成电路具有低成本和高效能的特性。这使得小型计算机得以发展，并且成为大型机的真正竞争者。格罗斯定理不再起作用，一组和一台大型机同样价格的小型机，和大型机相比，某些任务往往能处理得更快（尤其是那些可以并行处理的任务）。

从那时开始，小公司也有可能拥有它们自己的计算机。小型机可以完成类似设备控制或股票管理之类的功能。计算资源分布到整个企业标志着分布式计算概念的兴起。尽管如此，那时，一个组织里所有的计算机仍然独立地运行（图1-4）。

随着时间的推移，计算机用户的需求不断发展。用户不再满足于在一台计算机上独立地工作，有时候，他们需要与其他分支机构和办公室的同事交换计算机数据（通常是自动地）。为了满足这些需求，第一代局域网出现了（图1-5）。

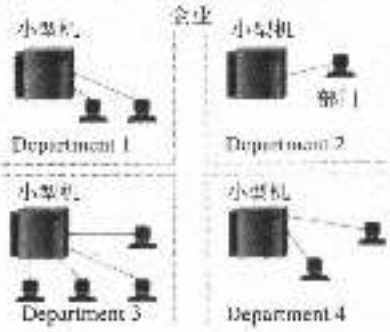


图1-4　同一家企业中几台独立的小型机

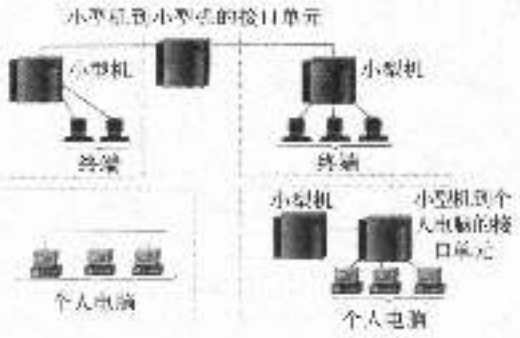


图1-5　第一代局域网的链路类型

局域网是一组聚集在一个相对较小范围内的计算机。虽然有些局域网可以被扩充到更大的范围（几十英里），但是，局域网通常覆盖的距离不超过1.5英里。大体而言，局域网是一个单纯连接的通信系统。

起先，人们使用非标准化的网络技术把计算机连接到网络上。

网络技术（network technology）是一整套协同工作的软件、硬件（例如，驱动器、网络适配器、电缆和连接器）以及在发送和接收比特数据时使用的规则。这些技术被用来构建计算机网络。

有些专用接口单元在通信链路上使用了专用的数据表示，使用了专用类型的电缆，它们只能将特定类型和型号的计算机连接起来，而且这些往往是它们自己设计的计算机。例如，PDP-11小型机到IBM360大型机的接口单元，或者是惠普小型机到LSI-11小型机的接口单元。

20世纪80年代中期开始，这一情况开始发生了变化。人们建立了将计算机连接到网络的**标准技术**（standard technology）。这些技术包括以太网、Arcnet、令牌环，以及后来的光纤分布式数据接口。

个人电脑的应用是这些技术发展的强大动力。个人电脑成为构建网络的理想元素。一方面，它们足够强大，可以支持网络软件。另一方面，它们也需要将它们自身的处理能力连接起来，用以解决复杂的任务，共享昂贵的外部设备和磁盘阵列。因此，个人电脑在局域网上非常盛行，不但是作为客户端，而且可以进行数据存储，完成一些处理中心的功能（例如，成为网络服务器）。随着个人电脑越来越流行，它们逐渐地淘汰了小型机和大型机。

所有的标准局域网技术建立在相同的交换原理上。这些交换原理已经在广域网的数据传输中大获成功，例如，分组交换原理。

建造局域网的过程由手工制作转向了使用网络技术的标准流程。为了构建一个网络，只需要根据具体的需求（例如，以太网）购买标准的电缆和网络适配器，用标准的连接器将适配器和电缆连接起来，然后在计算机上安装一种那时流行的网络操作系统（例如，Novell Netware）。

局域网的工程师发明了很多创新性的技术，这些技术极大地影响了用户工作的组织管理。类似于访问共享网络资源这样的任务得到了很大的简化。与广域网用户相反，局域网用户不再需要记忆复杂的共享资源的标识符。为了实现这一目标，系统用一种用户友好的方式列出所有可用的资源（例如，层次性的树型结构）。局域网的另一个优点是，当与远程的资源建立连接后，用户可以使用与访问本地资源相同的命令访问远程资源。大量的用户不再需要学习特殊的、同时也是相当复杂的网络命令。这不但是技术发展的结果，同时也成了技术发展的推动力。

所以，问题出现了：为什么仅仅当局域网出现后，用户才能享受这些便捷？这主要是因为局域网使用高质量的电缆。在局域网中，即使第一代的网络适配器也可以保证数据传输速率达到10Mb/s。由于局域网只覆盖有限的范围，这些线路的成本是可以控制的。也正是由于这一原因，节省带宽这一早期广域网技术的重要考虑因素，没有成为局域网协议发展的主要问题。在这一情况下，服务器间歇性地广播资源及服务，成为了组织局域网资源访问的主要机制。通过这些广播，客户机得到了可用的网络资源的列表，并把它们显示给用户。

20世纪90年代晚期，以太网成为了局域网无可争议的主导者。除了传统的以太网技术（10 Mb/s）以外，这一家族还包括快速以太网（100 Mb/s）和千兆以太网（1 000 Mb/s）。

简单的算法保证了以太网设备的低成本。数据传输速率的范围使得网络架构师在构建以太网时可以使用合理的方式，根据公司需求的特点，选择满足需求的以太网技术。从工作原理上说，所有的以太网技术都是非常相似的，这就简化了网络的维护和整合。

从时间顺序上说，表1-1列出了计算机网络发展历史的里程碑。

表1-1 计算机网络发展最重要的事件（按时间顺序）

计算机之间第一次全球连接。第一次批处理网络的实验	20世纪60年代晚期
通过电话网络传输数字语音	20世纪60年代晚期
出现大规模集成电路。第一台小型机。第一个专用标准的局域网	20世纪70年代早期
IBM系统网络体系结构	1974年
X.25技术标准化	1974年
第一台个人电脑的出现	20世纪80年代早期
现代模式的因特网的出现。在所有节点上安装TCP/IP协议栈	20世纪80年代早期
第一个标准局域网技术的出现	以太网——1980年
	令牌环网——1985年
	光纤分布式数据接口——1985年
因特网开始商业化运作	20世纪80年代晚期
万维网的发明	1991年

1.4 网络融合

1.4.1 LAN和WAN的融合

20世纪80年代晚期，局域网和广域网的区别变得非常明显：

- *通信链路的长度和质量*：局域网与广域网最明显的区别是，局域网网络节点之间的距离较短。这一因素使得网络开发者可以使用比广域网更高质量的通信链路。

- 数据传输方法的复杂性：由于物理通信信道的低可靠性，广域网要求更复杂的数据传输方法和更复杂的设备。
- 数据交换速率：在局域网中，数据交换速率（10Mb/s，16Mb/s和100 Mb/s）比广域网（从2.4 Kb/s到2 Mb/s）要高得多。
- 服务的多样性：局域网中的高数据传输速率允许网络开发者实现一系列的服务。这些服务包括访问和使用存储在其他联网计算机硬盘中的文件，共享打印机、调制解调器、传真机，访问集中式数据库以及电子邮件服务。广域网提供的服务主要局限于最简单形式的电子邮件和文件服务（对终端用户来说并不方便）。

渐渐地，局域网和广域网的区别开始消失。网络工程师开始通过广域网这一媒介，将独立的局域网连接起来。广域网和局域网的整合使它们的技术互相渗透。

数据传输方式的整合建立在基于光纤通信线的数字数据传输平台。这一传输媒介被几乎所有的局域网技术所使用，它的目的是为了实现110码距离以上的高速数据交换。几乎同时代的所有传输主干网都使用这一传输媒介，它提供了连接广域网设备的数字信道。

高质量的数字信道改变了广域网协议的要求。除了保证可靠性外，平均数据传输速率、优先处理对延迟高度敏感的分组等因素被带到了前端。这些改变在新的广域网技术中被反映出来，例如，帧中继和异步数据传输模式（ATM）。在这样的网络中，我们假设一比特的传输错误是一个极小概率的事件，如果发生，更为优选的做法是简单地丢弃整个分组。所有与分组丢失相关的问题都交由上层的软件模块来处理，这些软件模块并没有直接整合进帧中继网络和异步传输网络中。

因特网际协议的广泛采用使得局域网和广域网的整合成为可能。现在，这一协议被所有局域网和广域网技术所使用，包括以太网、令牌环、异步传输网络和帧中继网络。这更进一步创造了建立在不同子网基础上的统一的互联网[⊖]。

从20世纪90年代开始，工作在快速数字信道基础上的广域网极大地扩展了服务的种类。创建传输大量实时多媒体数据的服务成为可能，这些数据包括图像、视频和音频。万维网（WWW）作为一种超链接的服务，成为了因特网上的主要信息服务。它是一个很好的例子。这种服务的交互式能力超出了早期局域网提供的类似的服务。因此，局域网的构造者把这种服务从广域网中借鉴过来。这种将因特网技术移植到局域网的现象变得非常普遍，不久以后，专有名词**内部网（Intranet）**出现了。

现在，在局域网中就如在同广域网中一样，用户必须非常关注保护信息不被未授权访问。这是因为局域网已经不再独立了。局域网会不时地通过广域网的连接访问外部的世界。

最后，我们需要关注新技术的不断出现。最初，它们同时为两种网络提供服务。新一代技术的最好典范是异步传输模式[⊖]，由于它有效地将各种数据传输集中到一个传输网络中，它可以同时作为局域网和广域网的基础。从局域网发展出来的以太网技术是另一个例子。新的以太网10G标准允许数据以10Gb/s的速率传输，它可以被作为广域网和大规模局域网的主干网。

其他局域网与广域网整合的例子是**城域网（Metropolitan Area Network，MAN）**的出现，它

⊖ 互联网络是指通过路由器和其他设备连接在一起的一系列网络。通常，互联网络和一个单个网络同样工作。有时候它也被称为互联网（internet）。然而，不能把它和因特网（Internet）混为一谈。后者是指连接全世界成千上万网络的互联网络。

⊖ 异步传输模式是一种动态分配带宽的网络技术。异步传输模式使用固定大小的数据分组和两个数据传输点之间的固定的信道。异步传输模式被设计成支持多种服务，例如语音、图像、数据、视频。它允许电话和有线电视公司分配带宽给个人用户。

处于局域网和广域网之间。这些网络用来服务于大城市。

这些城域网使用数字传输信道，通常是光纤，主干网速度可以达到155Mb/s或更高。它们提供一种有效的方式将多个局域网连接，或者将局域网和广域网连接。起初，这些网络只是为了数据传输而设计的。现在，它们的服务范围被扩大了。例如，广域网支持视频会议和语音文字综合传输。现在的城域网具有多样化的服务，使得它们的客户可以将不同类型的通信设备连接在上面，包括用户交换机（PBX）。

1.4.2 计算机网络和电信网络的融合

计算机网络和电信网络融合的趋势每年都在增强。人们试图创建通用的**多服务网络**（multiservice network），这样的网络可以同时为计算机网络和电信网络提供服务。

电信网络包括电话、广播以及电视网络。它们与计算机网络最相似的地方在于它们都提供信息给终端用户。然而，这些网络以不同的方式提供信息。例如，计算机网络主要用来传输字符和数字信息，也可以称为数据。因此，计算机网络还有另外一个名称——**数据网络**（datanetwork）。电话和广播网络仅仅用来传输语音信息；电视网络既能传送语音也能传送视频。

尽管如此，计算机网络和电信网络的融合正在进行中。

首先，我们注意到提供给终端用户的服务类型正在被整合。首次尝试创建可以提供包括电话和数据传输服务的多服务网络催生了综合业务数字网（ISDN）技术。然而，实际上，现在的综合业务数字网主要提供电话服务。

现在看来，因特网是新一代全球多服务网络的主要候选者。它尤其吸引人的地方是它对多种传统服务的新的整合服务，例如，整合电子邮件、电话、传真、传呼等的信息服务。其中，IP电话服务最为成功，全世界上百万人直接或间接地使用它。尽管如此，因特网要真正成为真正新一代的网络，还有很长的路要走。

今天网络的技术整合建立在多种信息的数字传输、分组交换以及服务编程上。电话很久以前就开始向计算机网络整合。这之所以可行，是因为语音可以用数字化的方式表示出来，从而允许使用同样的数字信道传输电话和计算机数据。目前，电视网络也可以用数字方式传输信息。电话网络通常结合使用电路交换和分组交换。对于服务信息（也就是信号信息）的传输，它使用与计算机网络中相类似的分组交换协议；对于语音传输，它使用传统的电路交换方式。

电话网络提供的补充服务，例如通话转移、电话会议、电话投票等，可以通过使用**智能网**（Intelligent Network，IN）来实现。智能网是一种带有服务器的计算机网络，服务逻辑被编程在服务器中。

今天，即使在语音传输的领域中，分组交换方法也渐渐胜过了传统电话网络中的电路交换方法。这一趋势有其显而易见的原因：分组交换使通信信道和交换设备更有效地利用带宽。例如，电话通话中，有多达40%的总连接时间是处于停顿状态的。然而，只有分组交换才有能力去除这些停顿，并且使用释放出的信道带宽传输其他电话用户的数据。基于分组交换的因特网的流行，也成为分组交换更为优越的一个有利证据。

由于使用分组交换同时传输异构的数据（包括语音、视频、文字），开发新的方法保证**服务质量**（Quality of Service，QoS）变得尤为重要。对于实时信息，例如语音传输，服务质量的方法试图最小化传输延迟，同时，要保证平均数据传输速率和动态数据流量。

然而，我们不应该认为电路交换已经成为过时的技术，未来没有发展前途。处于当前这一技术发展的新阶段，它们也找到了它们自己的应用，但却是在其他的新技术里。

计算机网络也从电话网络和电视网络中借鉴了很多东西。因特网和公司内部网缺乏电话网络的高可靠性，计算机网络大量借鉴和使用了电话网络中常用的可靠性工具。

显而易见，下一代多服务网络并不是建立在某一种技术和方法的胜利之上。它将是融合的结果。这一融合过程将吸收每种技术的优点，把他们整合为一个混合体，来提供支持现有服务和新的服务所需的能力。为此，一个新的术语被引入——信息通信网络（infocommunications networks）。它阐述了当代网络的两个组成部分——（基于计算机的）信息和通信。由于这一新的术语还没有充分流行，我们将使用标准的、广为接受的术语——具有扩展含义的电信网络，包括计算机网络在内。

小结

- 计算机网络是计算机和通信技术发展的必然结果。它们代表了分布式计算机系统的一个特例，被认为是长距离传输信息的一种媒介。为了后一个目的，它实现了通信系统中所开发和采用的数据编码和多路复用技术。
- 根据地理范围，所有的网络可以分为以下类别：广域网（WAN）、局域网（LAN）和城域网（MAN）。
- 从时间顺序上说，广域网是最先出现的网络。它们将远隔几百英里的计算机连接起来。它们常常基于已有的、低质量的通信链路，因此数据传输速率较低。与局域网相比，广域网提供非常有限的服务，主要是文件传输和电子邮件，通常采用后台方式而不是实时方式。
- 局域网通常覆盖半径不大于1.5英里的范围。它们基于昂贵的、高质量的通信链路。相比于广域网，它们可以实现高速数据交换（大约100 Mb/s）。通常，局域网提供一系列的在线服务。
- 城域网服务于大城市。网络节点之间的距离相对较远（大约几十英里），它们也提供高质量的通信链路，支持高速数据交换。城域网提供了局域网的经济、高效的连接，并提供它们接入到广域网的机制。
- 计算机网络发展的最重要阶段是标准网络技术的出现，这包括以太网、光纤分布式数据接口以及令牌环。这些技术使得不同类型的计算机可以快速有效地接入网络。
- 在20世纪80年代晚期，局域网和广域网的显著区别是通信链路的长度和质量、数据传输的复杂性、数据交换速率、提供服务的覆盖范围以及规模。后来，随着局域网、广域网、城域网的整合，这些技术也进行了整合。
- 不同类型网络整合的趋势不限于局域网和广域网，也发展到了其他类型的电信网络，包括电话网、广播网和电视网。现在，研究重点集中在创建通用多服务网络，它能提供任何数据的有效传输，包括数据、语音和视频。

复习题

1. 多终端系统和计算机网络的不同之处是什么？
2. 何时第一次将计算机用长距离链路连接在一起，并取得了重要的成果？
3. 什么是ARPANET？
   a. 由属于美国军队和研究机构的超级计算机组成的网络
   b. 一个内部的科研网
   c. 组建广域网的技术
4. 第一个网络操作系统何时出现的？

5. 以下事件发生的顺序是什么？
   a. 万维网的发明
   b. 标准局域网技术的形成
   c. 语音开始以数字的形式通过电话网络传输
6. 哪些事件加速了局域网的发展？
7. 以下技术何时被标准化？以太网、令牌环和光纤分布式数据接口。
8. 列出计算机网络和电信网络融合的主要趋势。
9. 解释如下术语：多服务网络、信息通信网络和智能网络。

练习题

1. 解释为什么广域网的出现早于局域网。
2. 使用因特网查找相关的信息资源，寻找X.25技术与ARPANET网络的历史关系。
3. 你是否认为计算机网络的历史可以等同于因特网发展的历史？请阐述你的观点。

第2章 网络设计的一般原理

2.1 引言

当你开始学习具体的局域网、广域网或城域网，例如以太网、因特网协议（IP）或者异步传输网络（ATM）时，你很快就会发现它们具有许多共同点。然而，这些技术并非完全相同。恰恰相反，每种技术或协议都有其独特的特点。用户不能从孤立地掌握一技术领域的知识扩展到其他领域。提高学习效率的最有效的方法是从最适用的网络设计原理开始学习。这些原理构成了决定网络拓扑结构、路由方式、交换方式、信息流多路复用方式等的基础。因此，我们不须过多地解释那条众所周知的原理——“学习几条最基本的原理，可以帮助我们免去记忆过多的细节”。专家们一定非常熟悉各种细节，即便如此，如果能深刻理解设计的基本原理，依然会大有帮助。专家们可以有效地使用各种事实和细节，然后将这些事实和细节互相关联起来，构成一个和谐的系统。

建造数据网络的原理系统这一个过程看上去像是专门为了解决几个关键问题。这些问题中的大多数在各种电信网络中也非常常见。

在建造网络的过程中，交换是你将面对的最基础的问题之一。每一个传输活动中的网络节点都必须具备交换信息的能力（即，保证在多个网络用户之间的通信）。

使用网络选择传输信息流路由的原理直接影响了交换技术。选择**路由**（route）（在信息传输到目的节点的过程中，信息必须通过的网络节点的序列）必须同时实现两个目标。首先，每个用户的数据必须传输得越快越好，每个路由都需要有最小的延迟。其次，网络资源必须最有效地被加以利用，在任何时刻，对所有网络用户来说，网络必须传输最多的数据。最大的问题便是将这两个目标有机地统一起来（作为个人用户的个人中心主义的目标，和作为统一系统的整个网络整体的目标）。传统上，计算机网络并没有有效地解决这一问题，它们常常更倾向于解决单个的数据流问题，直到最近，它们才开始使用更为先进的路由方式。

在本章中，我们将介绍多路复用信息流的原理、共享传输媒体的原理、寻址的问题、选择网络拓扑结构的问题，以及逻辑和物理构建的问题。

2.2 共享计算机资源的问题

网络计算机最明显的优势之一，便是可以访问和使用连接到其他计算机的外部设备（磁盘、打印机、测绘仪等等）。与单个的计算机相似，网络计算机只可以直接管理那些物理上连到它们的设备。为了使不同计算机系统的用户可以共享外部设备，网络必须额外配置一些工具。我们来考虑最简单的网络，只有两台计算机（图2-1）。首先，我们考虑一台计算机和一个外部设备的交互作用。



图2-1 打印机共享

2.2.1 计算机与外部设备间的交互作用

为了组织一台计算机和一个外部设备之间的交互作用，这两者都必须配备外部的物理接口。

广义上说，接口是两个互相独立的通信对象之间正式定义的逻辑或物理边界。接口定义了参数、过程，以及两个对象之间的交互的特性。

物理接口（Physical Interface）（也被称做**端口**）是根据一套电路连接和信号特性来定义的。通常，它代表了与接点相连的连接器。每个接点都被赋予一个特定的目的。例如，有一组接点用来做数据传输，一个接点用来做数据同步，等等。一对插槽用一条电缆连接起来，电缆是一组连接两边相应接点的线路（见图2-2）。

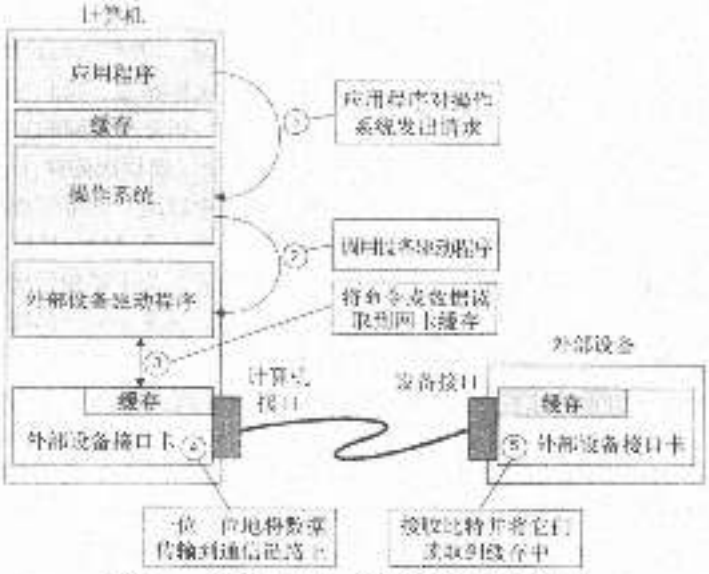


图2-2 一台计算机和一个外部设备之间的连接

逻辑接口（Logical Interface）是一组事先定义好格式的信息报文，以及一整套规定这种交换的逻辑规则。两个设备（在这个例子中，两个设备指一台计算机和一个外部设备）或程序使用这一共同的标准交换数据。

计算机使用两种典型的标准接口：并行接口（以字节为单位传输数据）和串行（以比特为单位传输数据）RS-232C接口（也被称为COM接口）。前者通常被用来连接打印机，后来使用得更为广泛，除了打印机之外，许多其他的设备也支持这种接口，包括猫绘仪，鼠标等。此外，还有其他连接特定外部设备的专用接口，例如一些用来从事物理实验的复杂设备。

计算机中的接口由硬件和软件共同实现。接口卡（一种硬件设备，也常被称为控制器或适配器）和计算机程序管理着物理设备。这些软件通常被称为外部设备驱动程序。

在外部设备中，接口常常使用**控制器**[⊖]（controller）这一硬件设备来实现。虽然也常常能遇见带有内置处理器的、由软件管理的控制器。这常见于具有复杂操作逻辑的外部设备中。一个很好的例子便是现在的打印机。

外部设备可以从计算机接收数据（例如，需要被打印的以字节为单位的信息）和指令。为了实现这一目的，外部设备控制器必须完成特定的操作。例如，打印机控制器可以支持一些简单的命令，例如，打印字符、换行、回车，从打印机中送出纸张，等等。打印机通过接口从计算机中获得这些命令，并且通过管理打印机的电子部件完成这些操作。

通常，通过接口进行的数据交换是双向的。例如，即使作为输出设备的打印机，也会返回状态信息给计算机。我们来考虑应用程序输出数据到打印机的操作顺序。

- 需要输出数据到打印机的应用程序请求操作系统完成一个输入输出操作，以下的数据必须

⊖ 接口卡、适配器、控制器这些术语常常被互换地回交叉使用。然而，为了区别安装在计算机内部的控制卡和打印机内置的控制器，对于前者，我们使用接口卡这一术语，对于后者，我们使用控制器。

在这个请求中指定：在随机存储器（RAM）中的数据地址、请求的外部设备的标识符，以及需要完成的操作。

- 接收到这个请求之后，操作系统调用应用程序指定的打印机驱动程序。在计算机部分，对于需要完成的输入输出操作而言，所有进一步的动作都由驱动程序控制的接口卡完成。
- 对于打印机驱动程序而言，它根据打印机控制器所能理解的命令工作：打印字符、换行、回车、从打印机中读出状态。驱动程序生成一系列命令编码，并将它们放到接口卡的缓存中，随后，缓存一个字节一个字节地将这些编码传送到打印机控制器。我们可以对同一个控制器开发不同的驱动程序，这些驱动程序使用相同的命令集，但使用不同的算法管理外部设备。
- 为使驱动程序和接口适配器协调工作，接口适配器实现了低层次的操作，这些操作允许它解释从驱动程序传送过来的数据和命令。这些数据和命令以统一的字节流的形式出现，接口适配器不需要去理解它们的含义。驱动程序接收到下一个字节之后，接口适配器开始按顺序地将比特传送到接口电缆中，每个比特用一个电信号代表。为了通知外部设备控制器即将开始传输下一个字节，接口卡在传输第一个比特的信息前，会先发出一个特定的开始信号。在传输完最后一个比特的信息后，接口卡会发出终止信号。这些开始和终止信号用来同步字节的传输。在识别出开始的比特后，控制器开始接收这些比特信息，并在接收缓存中组成一个字节。
- 除了信息比特之外，适配器还会传输奇验控制比特，这是为了保证数据交换的可靠性。在数据传输正确的基础上，控制器会解释这些接收到的字节，并且开始运行所请求的打印机操作。
- 在打印完文档的所有字符之后，打印机驱动程序通知操作系统它已经完成了这一请求，操作系统再将这一事件通知应用程序。

2.2.2　两个计算机间最简单的交互作用

让我们回到原来的问题：用户在计算机A上运行一些程序，他如何才能在连接到计算机B的打印机上打印出一些文本呢（图2-3）？

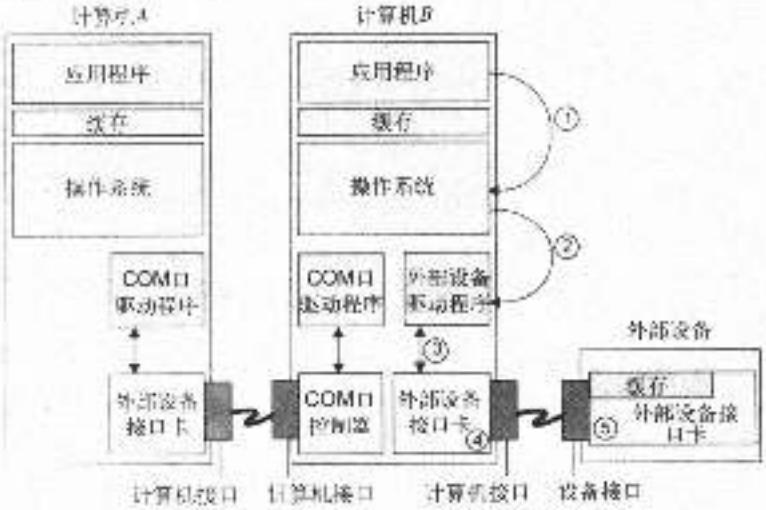


图2-3　打印机共享

运行在计算机A上的应用程序不能直接访问计算机B上的资源，例如磁盘、文件或者打印机。为了访问这些资源，应用程序可以呼叫那些运行在其他具有相应资源的计算机上的应用程序。这些请求使用报文（message）的形式来实现，并通过连接计算机的通信链路来传输。一般而言，这些报

文既包含指令（例如，打开一个文件），又包含需要处理的信息（例如，一个特定文件的内容）。

在很大程度上，连接在网络上的计算机之间进行交互的通信机制借鉴了计算机和外部设备间的通信方式。以最简单点的情况为例，可以通过使用一些工具来实现计算机之间的通信，这些工具对计算机和外部设备之间的通信起了组织作用。例如，可以通过串行接口——COM口来实现。双方的COM口都在各自驱动程序的控制下运行。在合作运行期间，它们保证了通过连接两台计算机的电缆，一个字节的信息可以从一方传输到另外一方。

说明 在真正的局域网中，有些功能可以通过网络接口卡（NIC）和它们的驱动程序来实现，前者有时也被称为网络适配器。从计算机的角度来看，网络适配器是一个正常的外部设备，和打印机控制器没有任何区别。

因此，我们可以定义两台计算机之间交换字节的机制。但是，这种最简单的工具对于解决我们的问题（用连接在另一台计算机上的打印机打印文本）并不够用。我们必须保证计算机*B*明白对于传输给它的数据，它应该完成何种操作，用哪一个设备来打印，被打印的文本是何种格式等等。应用程序*A*和应用程序*B*必须通过交换报文来解决所有这些问题。

此外，应用程序必须明白如何解释它们互相接收到的数据。为了达到这一目的，应用程序*A*和应用程序*B*的开发者必须在报文的格式和语义上达成一致。他们可以在以下方面达成一致：任何远程打印的执行必须以传输一个询问应用程序*B*是否已经准备完毕的报文来开始；这一报文必须包含计算机的标识符和发出请求的用户；特定的编码表示异常终止等等。我们马上就会看到，这些规则定义了应用程序之间交互的协议。

我们来考虑这个小型网络所有元素之间的交互，这些交互将允许运行在计算机*A*上的应用程序在连接在计算机*B*上的打印机上打印文档。

- 应用程序*A*必须生成一个报文给应用程序*B*，要求*B*打印一个文档。这一报文首先到达随机存储器的缓存中。为了将这一请求传输到远程计算机*B*上，应用程序*A*调用本地操作系统。本地操作系统调用COM口驱动程序，并将随机存储器缓存的地址传送给它，通过这一地址，可以找到所需要的报文。根据我们刚才介绍的方法，计算机*A*的COM口控制器和它的驱动程序与计算机*B*的COM口驱动程序和控制器交互，一个字节一个字节地传送报文给计算机*B*。
- 计算机*B*的COM口驱动程序始终等待外部来的信息。有些情况下，驱动程序被来自控制器的中断异步调用。驱动程序在接收到下一字节，并且检查完它的正确性后，将它读取到应用程序*B*的缓存中。
- 应用程序*B*接收报文、进行解释，根据报文的内容，应用程序*B*对本地操作系统发出一个请求，要求打印机完成特定的操作。计算机*B*上的操作系统将这一请求传送到打印机。
- 在打印的过程中，需要向应用程序*A*报告某些情况，在这个时候，我们使用了一个对称的设计。报文传输的请求从应用程序*B*到达计算机*B*上运行的本地操作系统。两边计算机上的COM口控制器和驱动程序组织报文一个字节一个字节地传输，然后，这一报文被读取到应用程序*A*的缓存中。

许多应用程序（文字或图像编辑器、数据库管理系统等等）的用户可能需要访问远程文件。显然，将刚才描述的应用程序*A*的功能建立到所有可能在网络环境下使用的标准应用程序中并不合理，虽然有些应用程序含有内置的网络功能。通常，这些应用程序对于数据交换的速度有相当严格的要求。对此，最有效的方法是开发特定的软件模块，这些模块被设计成专门向远程计算机发出请求，接收适用于所有应用程序的结果。这些软件模块通常被称为客户机和服务器。

客户机（client）是一种软件模块，它们从不同的应用程序生成请求报文，交给远程计算机，接收结果，并将这些结果传送回适当的应用程序。

服务器（server）是一种软件模块，它们始终侦听（listen）客户机请求。这些请求来自于连接到计算机上的网络或者特定的设备。服务器在接收到客户机的请求后，设法处理并完成这些请求，有时候本地操作系统也会参与其中。一台服务器可以串行或并行地完成多个客户机的请求。

客户机部件最方便最实用的特性便是它可以区分对本地的请求和对远程资源的请求。如果客户机程序可以做到这一点，那么，应用程序处理本地资源和远程资源的区别就不再重要了。因为客户机程序能识别出远程请求，并且将它们重定向（redirect）到远程机器上。因此，网络应用程序的客户机模块也被称作重定向器（redirector）。有时候，负责识别本地请求和远程请求的功能在一个单独的软件模块中实现。这时候，这一模块而不是整个客户机部件，被称为重定向器。

客户机和服务器软件允许计算机A上运行的所有应用程序对计算机B上资源（打印机、文件、传真机等）的远程访问。为了使计算机B上运行的应用程序能够访问计算机A上的资源，这一设计必须对称地使计算机B配备客户机软件，计算机A具备服务器模块。

图2-4描述了应用程序和本地操作系统与客户机和服务器交互的方式。虽然我们只考虑了两台计算机之间最简单的交互方式，但是，访问远程打印机的程序功能，与运行在具有许多计算机和复杂连接方式的网络上的操作系统具有许多相似之处。

说明　术语客户机和服务器指软件模块和计算机这一整体。如果一台计算机向网络上的其他计算机提供资源，则它被称做服务器。使用服务器提供资源的计算机称做客户机。有时候，同一台计算机可以同时是客户机和服务器。

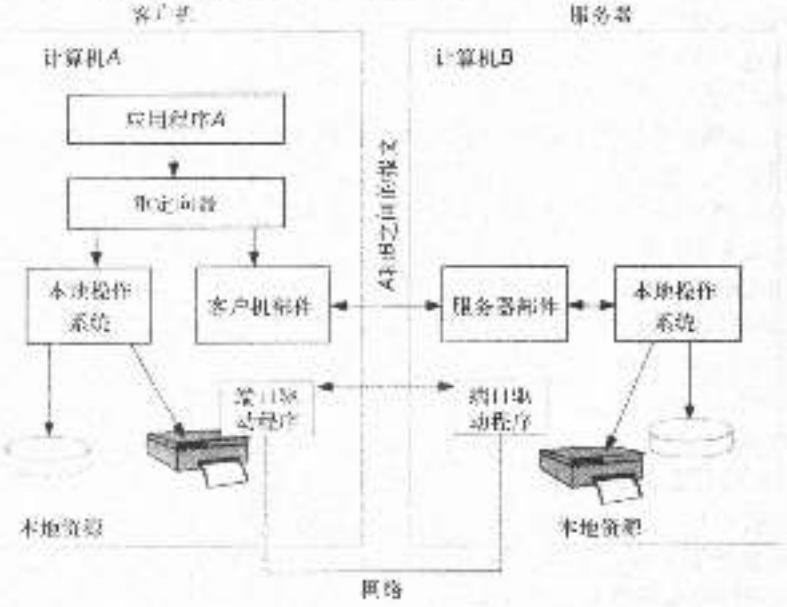


图2-4　连接两台计算机时，软件部件之间的交互作用

2.2.3　网络应用程序

向用户提供特殊类型的资源的共享访问，诸如文件，也被称做提供服务（在这个例子中，称做文件服务）。通常，网络操作系统给它的用户提供多种类型的网络服务——文件服务、打印服务、电子邮件服务、远程访问服务（RAS）等。实现网络服务的程序被归类为分布式程序。

分布式程序由多个互相交互的部件（在图2-5的例子中，为两个文件的软件部件）组成。通常，每个部件可以在不同的网络计算机上运行。

网络服务是分布式系统程序，常常是操作系统的一部分。



图2-5 分布式应用程序模块之间的交互

然而，还有分布式的用户程序——应用程序（application）。分布式（distributed）应用程序也由几个部件组成，每个部件执行特定的操作完成特定的用户任务。例如，这种应用程序的一部分运行在用户的工作站上，支持特定的图形化用户界面（graphic user interface, GUI）；另一个模块运行在一台高性能的专用计算机上，统计处理用户提供的数据；第三个部分可能将结果存储到一台安装有标准数据库管理系统（DataBase Management System, DBMS）的计算机上。分布式应用程序充分地运用了数据网络提供的潜在的分布式处理能力。因此，它们通常被称做网络应用程序（network application）。

说明 需要指出的是，不是所有运行在网络环境下的应用程序都可以被归类为"真正的网络应用程序"。有许多非常流行的应用程序，虽然它们可以在网络环境下运行，但并不代表它们是分布式程序。这是因为它们的部件不能分布到不同的计算机上。然而，由于操作系统中内置的网络服务，即便是这些应用程序也可以得益于网络提供的优势。局域网的很大一部分历史与这些应用程序的使用联系在一起，让我们来考虑用户如何在dBase数据库管理系统下工作。这一系统在那时非常流行。事实上，dBase是PC机上最早的数据库管理系统之一。通常，所有网络用户访问的数据库文件都存放在一台文件服务器上。数据库管理系统作为一个单独的软件模块安装在每个客户机上。起初，dBase只是用来处理本地数据（例如，数据与数据库管理系统软件位于同一台计算机上）。用户在本地计算机上启动dBase，程序在本地硬盘上搜索数据，并不考虑网络的存在。为了使用dBase处理远程数据，用户必须访问文件服务。这些服务将数据从服务器传送到客户端计算机，并且使数据库管理系统觉得这些数据仿佛是存储在本地。

在20世纪80年代中期，局域网上的大部分应用程序并不是真正的分布式程序。这不难理解，因为这些程序最早是为单机编写的。随着网络越来越流行，这些应用程序被安装在网络环境下。虽然开发真正的分布式应用程序会带来很多的好处（例如，减少网络流量，根据计算机的角色专门化），但是，事实上，这是一个很难的任务。我们还需要解决许多其他问题。开发人员需要决定应用程序需要多少个模块，这些模块如何相互作用，以保证某些模块失效或者发生错误后，其他模块可以正确地终止，等等。即使在现在，现有的应用程序也只有一小部分是真正分布式的。然而，很明显，未来是属于这种类型的应用程序的，因为它们可以充分应用网络在并行数据处理方面的潜力。

2.3 使用通信链路的物理数据传输的问题

即使只考虑连接两台计算机的最简单的网络，我们都可以看到许多适用于任何计算机网络的

问题。首先，我们需要分析使用通信链路进行信号物理传输的问题。

2.3.1 编码

在计算时，信息是以二进制编码的形式表示的。在计算机内部，离散的电信号代表1和0。

以电或光信号形式表示的数据称为编码。

有许多方法可以对二进制数据进行编码。例如，使用所谓的电位方法，特定的电压等级代表1，另一个电压等级代表0。另一个可能的方法是脉冲方法，在这一方法中，不同极性的脉冲被用来代表二进制数据。

类似的数据编码方法也可以被用在计算机和计算机之间使用通信链路进行的数据传输。然而，这些通信链路和计算机内部的通信链路在许多特性上并不相同。外部通信链路和内部通信链路最显著的区别，在于外部通信链路要长得多。此外，它们在起屏蔽作用的计算机机箱之外，经常会经过有外部电磁干扰的地方。所有这些因素产生了矩形脉冲的干扰（例如，脉冲前端的扭曲），这些干扰比机箱内部的干扰要严重得多。因此，为了保证通信链路的接收端可以可靠地识别脉冲，在计算机内部和外部，并不总是使用相同的传输速率和编码方式。例如，由于通信链路的高负载，脉冲前端上升的比较缓慢。因此，为了避免两个相邻脉冲前沿和后沿的重合，并且保证脉冲有足够的时间上升到所需的等级，我们需要降低传输速率。

在计算机网络里，离散数据的脉冲编码和电位编码方法都在使用，此外还有一种**调制**（modulation）方法——一种在计算机内部从不使用的特定的数据表示方法（图2-6）。当使用调制时，离散信息用频率的正弦信号来表示，并且通过可用的通信链路可靠地传输。

脉冲编码和电位编码在高质量的通信链路（*high-quality communication link*）中使用，基于正弦信号的调制更适合于链路在信号传输过程中可能会受到显著干扰时使用。例如，在使用电话线传输数据的广域网中，常常使用调制方法，这些电话线被用来以模拟的方式传输语音，因此，很难适合直接传输脉冲。

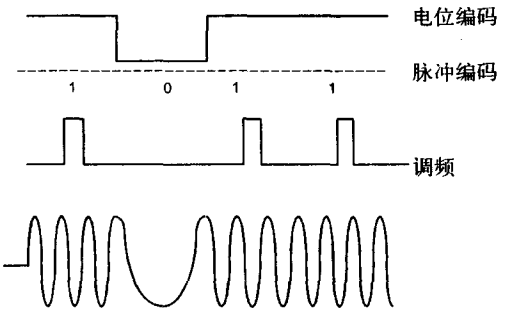


图2-6　表示离散信息方法的例子

连接计算机的通信链路中的线路的数量（*the number of wires*）也影响着信号传输的方式。为了减少网络通信链路的成本，最常用的做法是减少线路的数量。为此，常常使用顺序的、一个比特一个比特的数据传输，这一方法只需要一对线路。在计算机内部的数据传输同时能够传输组成一个字节的所有比特，有时候甚至是几个字节。

计算机和计算机间传输信号所面临的另一个问题是一台计算机的发送端和另一台计算机的接收端之间的同步问题。在一台计算机内部组织硬件模块的交互时，这一问题可以很容易地解决，因为所有的模块都可以根据时钟脉冲发生器来进行同步。计算机间通信同步的问题可以用两种方法来解决。一种是使用单独的线路交换特殊的同步时钟脉冲，另一种是使用预定义码或与数据脉冲不同的脉冲编码来进行周期性的同步。

即使选择了恰当的数据交换速率，根据需要的特性使用了通信链路，选择了发送端和接收端的同步方法，传送数据中有些比特受到干扰的可能性依然存在。为了保证计算机之间数据传送的可靠性，常常使用一个标准化的方法——计算**校验和**（checksum），并且在每个字节或者若干个字节之后，使用通信链路传送这一校验和。通常，数据交换协议包含必备的收到确认的部分。这一**确认**（acknowledgement）由接收端发送给发送端，用以确认数据接收的正确性。

2.3.2 物理链路的特性

使用物理链路传输数据有许多重要的特性。这里，我们将介绍一些你现在就需要了解的特性。在第6章中，我们会介绍一些其他的特性。

- **网络负载**（offered load）是从用户到网络输入端的数据流。网络负载以数据输入到网络的速率为特征。通常，传输速率以几个比特每秒、千比特每秒、兆比特每秒等等来衡量，表示为b/s、Kb/s、Mb/s等等。
- **信息速率**（information rate）或**吞吐量**（throughput）（这两个术语是同义词，可以相互替代）是数据流经过网络的数据速率。它可以比网络负载低，因为网络就像其他系统一样，会工作在用户并不希望的状态下。例如，数据可能会发生错误或者丢失，因此，实际信息速率可能会降低。
- **容量**（capacity）被定义为使用一种特定类型链路的最大可能传输速率。其典型的特征在于它的数值同时取决于媒介的物理特性，以及使用这种媒介后，选择的传输离散信息的方法。例如，光纤以太网的链路容量是10Mb/s。这一数值反映了特定介质（在这一例子中，指光纤）和所选技术（以太网）结合体的速率极限。这一速率极限取决于数据编码方法、信息信号的时钟频率，以及其他的参数。对于同样的媒介，可以开发另一种数据传输技术，形成另一个容量值。例如，快速以太网使用相同的光纤连接，但最大传输速率可以达到100Mb/s；千兆以太网将最大传输速率提高到1000Mb/s。通信设备的发送端必须工作在与链路容量相同的速率。这一速率有时候被称做发送端的比特率。
- **带宽**（bandwidth）这一术语的使用有时候容易使人误解，因为它有两种不同的含义。首先，它可以用来指代传输介质的物理特性。这时，这一术语是指频带的宽度，在这种频带宽度下，通信线路传输数据时不会发生明显的传输错误。这一术语的起源来自于这一定义。另一方面，这一术语也可以用来表示容量。当表示频带宽度时，带宽的单位是赫兹（Hz），当表示容量时，带宽的单位是比特每秒。这一术语的意思必须根据上下文来确定，虽然有时候这并不容易。当然，对于不同的特性使用不同的术语会更好，但是，有些传统很难更改。带宽这一术语的双重意义已经变得非常常见，在许多标准或者书籍中都会遇到。因此，我们也使用这种方式。除此之外，我们考虑到这一术语的第二种意思更常用一些，因此，除非和实际意义不符，我们更多地使用这一术语的这层含义。

通信链路的下一组特性与*单向和双向*（*one or both directions*）传输数据的可行性相关。

在两台计算机的交互过程中，常常需要双向传输信息——从计算机*A*到计算机*B*，并且反过来。即使有时候，用户感觉他们仅仅是接收信息（从互联网上下载音乐）或是发送信息（发送电子邮件），但实际上，信息的交换是双向的。有两种数据流——对用户有实际意义的主数据流，以及反向传输的辅助数据流。这一辅助数据流由对主数据流的确认所组成。

物理链路根据它们双向的信息传输容量来分类。

- **双工链路**（duplex link）可以在两个方向上同时传输信息。双工链路可以由两条物理介质组成，每一条在一个单一的方向上传输信息。也可以使用同一介质同时双向传输数据。然而，在这种情况下，我们需要使用额外的方法来分开每一个数据流。
- **半双工链路**（half-duplex link）也可以在两个方向上传输数据。这一传输并不是同时的，相反，它是轮流的。在特定的时间段，信息以一个方向传输，在下一个时间段，信息以相反的方向传输。
- **单工链路**（simplex link）只允许单项的信息传输。通常，双工链路由两个单工链路构成。

本书的第二部分“*物理层技术*”将更详细地介绍物理数据传输的各个方面。

2.4　多台计算机交互的问题

到目前为止，我们介绍了只有两台机器的最简单的网络，当更多的计算机加入网络后，新的问题出现了。

2.4.1　物理链路的拓扑

随着多于两台计算机互连这一问题的出现，我们需要决定如何去连接它们。换句话说，必须选择物理链路的配置，也称为*拓扑结构*（*topology*）。

网络拓扑结构（network topology）指一个图的形状，图的顶点代表网络节点（例如计算机）和通信设备（例如路由器），图的边代表它们之间的物理连接或者信息连接。

随着需要被连接的设备的增加，可能的配置的数量也会迅速增加。例如，可以使用两种方法连接三台计算机（图2-7a）。对于有四台计算机的配置，在拓扑上就有六种不同的配置（假设所有的计算机都相同），如图2-7b所示。

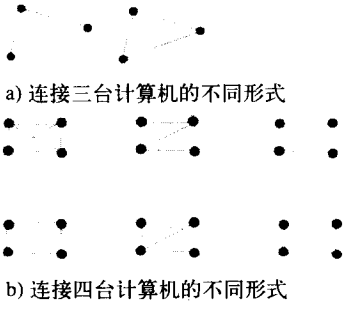


图2-7　连接多台计算机进入网络的可能的形式

每一台计算机都可以连接到其他所有计算机，或者所有计算机可以被顺序连接。在后一种情况下，假设它们通过传送“中转”报文来通信。在这个例子中，**转发节点**（transit node）必须具备特殊的工具，使它们可以执行这一特定的中转操作。通用计算机和特殊设备都可以作为转发节点。

大多数的网络特性取决于拓扑结构的选择。例如，节点之间有多条路由可以增加网络的可靠性，并确保传输链路的负载平衡。有些拓扑结构可以很方便地连接新的设备，这使得网络具有可扩展性。如果出于经济的考虑，往往选择通信链路总长度最小的拓扑结构。

在大量可能的配置中，我们需要区分全连接和部分连接的拓扑结构。

全连接拓扑结构（fully connected topology）（图2-8a）相当于网络中的每一台计算机都和其他所有的计算机直接相连。虽然它在逻辑上非常简单，但是，这种拓扑结构体积庞大、效率不高。每一台计算机都必须具有大量的通信端口，这些通信端口要足够连接网络上其他所有的计算机。每两台计算机之间必须有物理链路。（如果一条链路不能双向传输，那么必须具有两条链路。）在大规模网络中，全连接拓扑结构很少被使用，因为假如要连接N个节点，则必须配备$N(N-1)/2$条双向物理链路（即链路的数量与节点的数量平方相关）。通常，只有在几台计算机间或者是连接很少计算机的小规模网络里才使用这种拓扑结构。

其他所有类型的网络都基于**部分连接拓扑结构**（partially connected topology），当两台计算机之间交换数据时，可能要使用到其他的网络节点作为转发节点。

网型拓扑（mesh topology）[⊖]从全连接拓扑结构而来，它删除了其中的一些链路（图2-8b）。网型拓扑结构可以把大量的计算机连接起来，适合于大规模的网络。

在**环型拓扑**（ring topology）的网络里（图2-8f），数据通过计算机和计算机之间的环传输。环最大的好处是它可以提供冗余的链路。每一对节点都由两条路由连接——顺时针和逆时针。环这一配置非常方便提供反馈，因为数据经过整个环后，会回到它的起始节点。因此，起始节点可以控制数据递送到目的节点的过程。通常，环的这一特性被用来测试网络的连通性，以及用来搜索哪些节点不能正常工作。另一方面，在环型拓扑的网络中，我们也需要采取特殊的措施来保证当一台计算机失效或者是暂时失效时，其他所有网络节点的通信连接还能正常工作。

⊖ 有时候，术语网状（mesh）也被用来指代全连接拓扑或者是与它非常相近的拓扑结构。

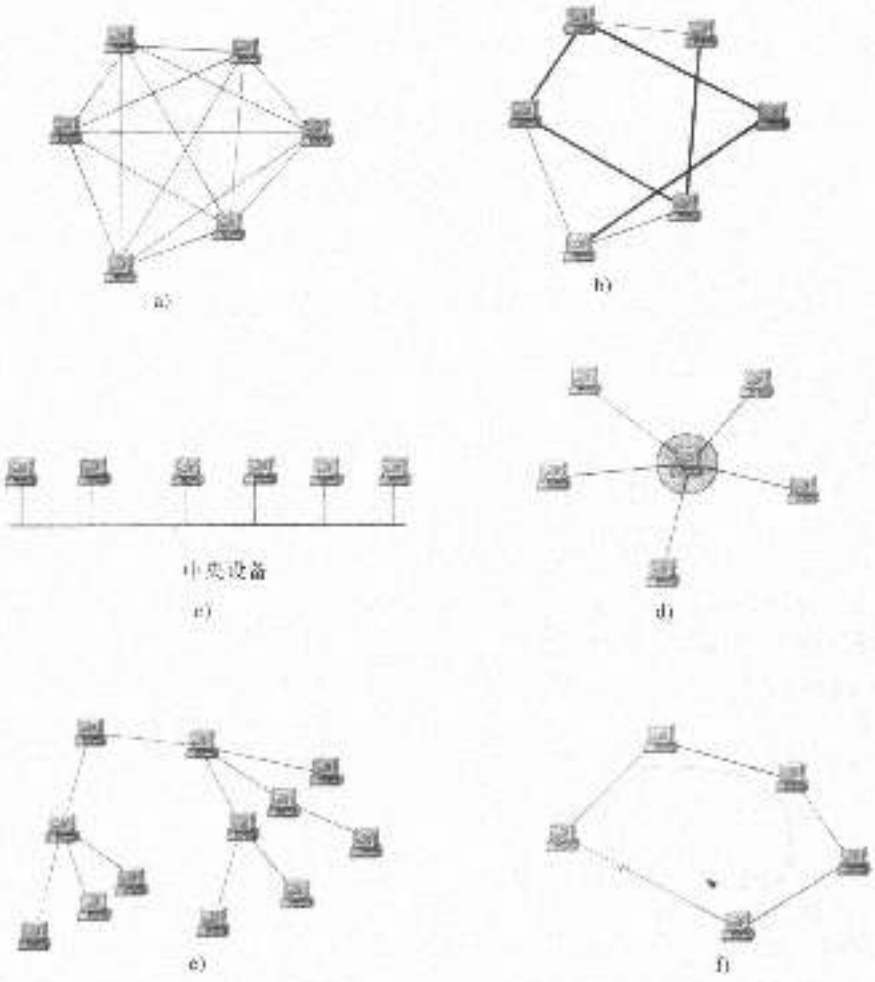


图2-8 典型的网络拓扑结构

星型拓扑结构（star topology）（图2-8d）假设每台计算机直接连接到一个叫做**集中器**（concentrator）[⊖]的中央设备上。集中器的功能包括将信息从一台计算机重定向到另一台特定的计算机，或者重定向到所有组成网络的计算机上。可以使用通用的计算机作为集中器，或者使用专门的网络设备。这一网络拓扑结构有一些缺点，例如，网络设备的成本较高，因为需要购买一些专门的中央设备。除此之外，集中器可用端口的数量限制了可增加的网络节点的数量。有时候，我们需要使用多个集中器来构造网络，使用层级星型把一个接在另一个之上（图2-8e），所产生的结构称为**层级星型**（hierarchical star），或者**树型**（tree）。目前，树型结构是局域网和广域网中最常见、最广泛使用的结构。

公共总线型（common bus）配置是星型拓扑的一种特殊形式（图2-8c）。此时，中央设备的角色委派给一条无源电缆，若干台计算机根据线路或者设计连接到这一电缆上。大多数无线网络也具有相同的拓扑结构。然而，在这种情况下，常见的无线传输介质担任公共总线的角色。信息通过电缆来传输，所有连接在这一电缆上的计算机可以同时获得这些信息。这一设计最大的优点

⊖ 在这个例子中，集中器这一术语是一个广义的用法，指任何具备多个输入并且可以作为一个中央元素（例如，可以使用一个交换机或者路由器）的设备。

是低成本和连接新节点的简便性。公共总线型最大的缺点是它的低可靠性，因为电缆或者这么多连接器中的任何缺陷都可能使整个网络瘫痪。公共总线型的另一个缺陷是它的低效率，因为这种连接方式意味着在某一时刻，网络上只能有一台计算机传输数据。因此，通信链路的带宽总是被许多网络节点平分。直到最近，公共总线型还是局域网最常见的拓扑结构之一。

小型网络总是具有典型的拓扑结构之一——星型、环型或者公共总线型，大型网络却有许多计算机之间任意的连接。在这样的网络中，我们可以分辨出具有典型拓扑结构的任意连接中的一部分（子网）。因此，这些网络被称为**混合拓扑结构**（mixed topology）的网络（图2-9）。

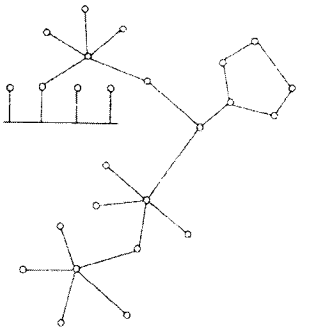


图2-9　混合网络拓扑结构

2.4.2　网络节点的编址

当连接三台以上的计算机时，我们就需要考虑编址的问题。准确地说，这是网络接口编址的问题[⊖]。一台计算机可以有多个网络接口。例如，在一个物理环中，每一台计算机必须至少配备两个网络接口，来保证它可以和它的邻居相连。为了创建一个具有N台计算机的全连接的结构，每台计算机必须配备$N-1$个网络接口。

根据编址接口的数量，网络地址可以分成如下几类：

- **单播地址**（unicast address）用来标识单个的接口。
- **多播地址**（multicast address）标识几个接口组成的组。被标记为多播地址的数据会被递送到属于那个组的每一个节点。
- 许多网络技术支持所谓的**广播地址**（broadcast address）。具有这些地址的数据必须被递送到所有的网络节点。
- IPv6这一新的因特网协议定义了一种新的地址——**任播地址**（anycast address）。与多播地址相似，任播地址定义了特定的地址组。然而，发送到这种类型地址的数据必须被递送到组内的任意地址，而不是属于这一组的所有地址。

地址可以是**数值**（numeric）（129.26.255.255），也可以是**字符**（symbolic）（site.domain.com）。

字符地址（名字）的目的是为了用一种易于被人们接受的方式标识网络节点；因此，它们通常和语义相联系。字符地址更容易记住。大型网络中使用的网络名字有层次型的结构，例如ftp-arch1.ucl.ac.uk。这一有意义的名字表示，被分配给这一地址的计算机支持伦敦大学学院（University College London，或ucl）网络中的ftp-archive，并且，这一网络与英国（uk）的教育网分支（ac）相关。在伦敦大学的网络内部工作时，这么长的字符名称是冗余的。简化的名字（ftp-arch1）比全名更方便。

对人类用户来说，字符名称更方便。然而，由于它们多变的格式，以及它们的长度可能很长，它们在网络上的传输是低效的。

*在一种特定的编址方式下，所有有效的地址的全集称为*__地址空间__（address space）。

地址空间可以有**扁平**（flat）（线性，参见图2-10）结构或者**层级**（hierarchical）（图2-11）结构。

对于前者而言，地址空间是没有结构的。一个典型的扁平数值地址的例子便是所谓的**MAC地址**（MAC address），它用作局域网中网络接口的唯一标识。这类地址通常供硬件使用。因此，它们尽可能的短。约定俗成，MAC地址使用二进制或十六进制（例如，0081005e24a8）的格式书写。MAC地址不需要人工指定，因为它们通常由硬件制造商硬编码。出于这一原因，MAC地址也被称

⊖ 有时候，我们会使用简化的术语“网络节点地址”，而不是精确的术语“网络接口地址”。

做**硬件地址**（hardware address）。使用扁平地址并不是一种灵活的方案，因为当网络硬件更换后（例如，网络适配器），网络接口的地址也变更了。

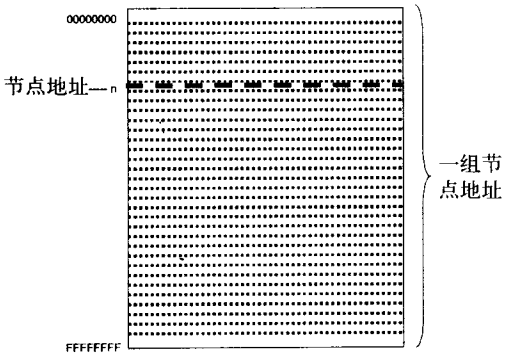


图2-10　地址空间的扁平结构

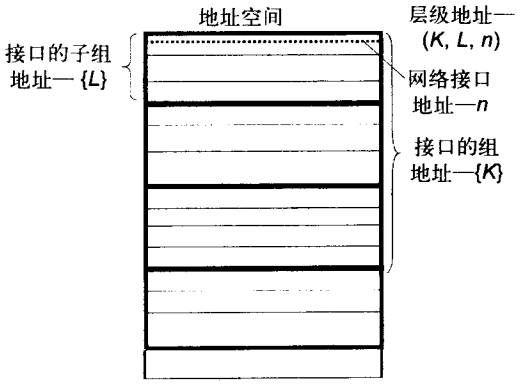


图2-11　地址空间的层级结构

当使用**层级编址**（hierarchical addressing）方法时（图2-11），终端节点的地址便由以下三个部分决定：标识特定节点属于哪个组的**组标识符**（group identifier）（K）；**子组标识符**（subgroup identifier）（L）；唯一标识子组中节点的**节点标识符**（node identifier）（n）。在大多数情况下，层级地址比扁平地址更有效。在具有成千上万节点的大型网络中，使用扁平地址会产生巨大的额外负担，因为终端节点和通信设备必须与具有成千上万条记录的地址表共同工作。相反地，数据传输中的层级地址方法允许在某一时刻前，只使用地址的最左边的部分（例如，K），然后，为了进一步缩小地址范围，再使用下一部分（L），最后才使用最右边的部分（n）。

IP和IPX地址是层级数值地址的典型代表。它们支持两层的层级结构，此时，一个地址被划分为高有效部分——网络数值，和低有效部分——节点数值。这一划分允许网络间报文的传输基于网络数值。只有在报文被递送到相应的网络中后，才会使用节点数值。这一方法和邮件递送中使用的街道地址相似。只有在信件被递送到目标城镇之后，才会使用街道名。

实际上，几种编址方式同时在使用。因此，计算机的网络接口可以有多于一个的地址（或名称）。每一个地址在方便的时候被使用。为了将地址从一种形式转换到另一种形式，人们使用称为**地址解析协议**（address resolution protocol）的特殊协议。

用户根据层级的字符名称为计算机编址。在网络上传输的报文中，这些字符名称被自动替换为层级的数值地址。通过使用这些数值地址，报文从一个网络传输到另一个网络，当报文被递送到目标网络后，层级的数值地址不再被使用，而使用计算机的扁平硬件地址。

对于解决不同类型地址之间对应关系的问题，既可以使用集中式的工具，也可以使用分布式的工具。

对于集中式的方法，网络中有专门的计算机，称为**名称服务器**（name server），它们存放不同类型名称（例如，字符名称和数值地址）的映射关系表。所有其他的计算机请求名称服务器根据一台计算机的字符名称确定它的数值标识符。

当使用分布式的方法时，每一台计算机自行存放所有分配给它自己的各类地址。需要根据已知的层级地址确定扁平硬件地址的计算机向网络送出一个广播请求。所有网络上的计算机将这一广播报文中的层级地址和它们自己的地址相比较。地址相符的计算机发送一个回答报文，报文中包含所需的扁平硬件地址。TCP/IP协议栈的**地址解析协议**（Address Resolution Protocol, ARP）使用了这一方法。

分布式方法的好处是它不需要一台专用的计算机，通常，这台计算机还需要人工配置地址映

射表。然而，分布式方法也有它的不足，其中最主要的是它需要发送广播报文，这些报文会给网络造成负担，因为它们被发送给所有的节点。因此，分布式方法只在小型的局域网中使用。对于大型的网络，集中式的方法更为常见。

到目前为止，我们集中讨论了网络节点（例如，计算机或特殊的通信设备）中网络接口的地址。然而，网络上代表数据传送目的端的并不是计算机或路由器，而是在这些设备上运行的软件。出于这一原因，目的地址除了标识目的设备接口的信息外，还必须标识数据要到达的进程的地址。当数据到达目的地址标识的网络接口时，计算机上运行的软件必须将数据转送到相应的程序。显然，程序的地址不需要在整个网络上保持唯一，只要保证它在一台计算机上唯一就足够了。TCP/IP协议栈中使用的TCP和UDP端口号（port number）代表了程序地址的例子。

2.4.3　交换

假设计算机已经根据一个特定的拓扑结构在物理上连接起来，并且已经选择了一种特定的编址方法，现在，最重要的问题是要解决：网络中节点之间传输数据需要使用什么方法？当使用部分连接的网络拓扑结构时，该问题变得极其复杂。在这一情况下，在任选的一对节点（用户）之间的数据传输，总是需要经过一些转发节点。

通过网络转发节点连接终端用户的过程称为**交换**（switching）。从源节点到目的节点路径上的节点的序列称为**路由**（route）。

例如，在如图2-12所示的网络中，节点2和节点4没有被直接地连接起来，它们之间传输数据必须通过转发节点（例如，节点1和节点5）。节点1必须将数据从接口*A*传输到接口*B*，节点5必须采取同样的行动，将数据从接口*F*传输到接口*B*。在这个例子中，路由可以描述为：2-1-5-4，其中节点2是源节点，节点1和节点5是转发节点，节点4是目的节点。

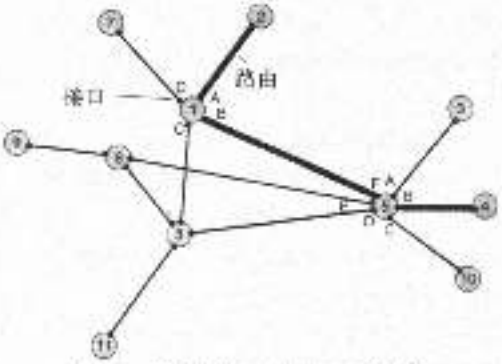


图2-12　通过网络转发节点进行交换

2.5　通用的交换问题

就最常见的形式而言，交换可以由以下一组相互关联的任务来描述：

- 确定信息流，为此，必须定义路由
- 发送信息流
- 流转发（例如，在每个转发节点上，运行流识别和本地交换）
- 流的多路复用和解多路复用

2.5.1　流定义

显然，通过一个转发节点可以建立多条路由。例如，所有节点4（图2-12）发出的数据必须经过节点5，同时，所有要发送到节点3、4、10的数据也必须经过节点5。转发节点必须具备识别到来的数据流的能力，并将每一数据流转发到正确的接口，这一接口可以将数据传送到所要求的目的节点。

信息流（*information flow*）或数据流是选择的数据序列，它带有一组公共属性。这些属性可以将它与所有的网络数据区分开来。

例如，所有来自一台特定计算机的数据可以被确定为一个单独的流，源地址可以作为统一的属性。同样的数据可以被表示为一系列更小的子流，每一个使用目的地址作为区分彼此的属性。最后，每一个这样的子流还可以被进一步划分成不同的网络应用程序产生的数据流。例如，电子

邮件、文件复制程序或Web服务器。

构成流的数据可以被表示为数据单元——**分组**（packet）、**帧**（frame），或**信元**（cell）。

注意　除了data flow这一数据流的概念之外，还有另一个称为data stream的数据流的概念。通常，data flow的速率不平均，而data stream有一个恒定的速率。例如，通过因特网传输一个网页时，流量负载用data flow来表示；在因特网音乐广播中，那便是一个data stream。对数据网络而言，不平均的速率更为典型；因此，在大多数情况下，我们使用流（flow）这一术语。另一术语stream则仅仅当需要强调这一过程的等时性时才会使用。

在交换数据的过程中，目的地址是一个必需的属性。基于这一属性，到转发节点的整个数据流被划分为多个子流，每一个子流被转发到相应的数据转发路由的接口上。

对每一对终端节点而言，源地址和目的地址确定了一个数据流。通常，也可以将两个终端节点之间的数据流表示为一系列子流，每一子流在其特定的路由上传输。同一对终端节点可以执行多个使用网络交互的应用程序。同时，每一个应用程序可能有其自己对网络的需求。在这种情况下，路由选择必须根据应用程序的需求来决定。例如，对于文件服务器来说，将大量的数据转发到一个高带宽的通信链路是非常重要的。对于发送必须立即处理的短消息的管理系统来说，最重要的是连接链路的可靠性，以及所选路由的最小延迟程度。除此以外，即使对于对网络有相似需求的数据而言，也有可能需要建立多条路由，通过并行使用不同的通信链路来加快数据的处理。

流的属性可以是*全局的*（*global*），也可以是*本地的*（*local*）。对于前者，它们在整个网络范围内唯一地标识了流。对于后者，它们只作用于特定的转发节点。例如，使用全局属性标识流的一个典型例子是一对终端节点的唯一地址。在一个特定的设备中局部定义的流的属性，可以是特定转发节点的接口的标识符，数据将通过这一接口传输。为了更好地描述这些定义，让我们回到图2-12中的网络配置。在这个例子中，节点1可以被配置为传输所有从接口*A*到接口*B*的数据，以及所有从接口*D*到接口*C*的数据。这一规则可以将从节点2到来的数据流和从节点7到来的数据流分开；这也允许它们通过不同的网络节点进行中继传输。在这个例子中，来自节点2的数据将通过节点5，来自节点7的数据将通过节点8。

还有另外一种特定的流的属性称为*流标记*（*flow label*）。流标记是流中的所有数据带有的一个特定的数字。该标记有一个全局唯一的值，在整个网络中唯一地标识该流。在这种情况下，它被分配给流的数据单元，在从源节点到目的节点的整个路由上永不变更。在有些情况下，本地流标记可以在节点到节点的传输过程中，动态地改变它们的值。

因此，在交换过程中的流识别是基于属性的，这些属性除了必备的目的地址外，还可以包含诸如特定应用程序标识符等信息。

2.5.2　路由

路由的问题包含两个任务：

- 确定路由
- 将选定的路由通知网络

解决数据传输的路由选择问题，包括确定转发节点的序列和它们的接口，通过这些转发节点和接口，数据可以被递送到目的地址。确定路由是一个复杂的任务，特别是在网络配置允许在一组交互的网络接口之间有多条路由时。最常见的，人们往往根据特定的标准只选择一条路由，称为*最优路由*（*optimal*）[⊖]。这些标准可以用来作为最佳标准，包括额定带宽、通信链路负载、特

⊖　事实上，为了减少计算量，从数学的角度上说，人们往往选择一条非常接近于最优路由的路由，而不是真正的最优路由。

定链路的延迟、转发节点的数量以及通信链路或转发节点的可靠性。

即使当两个终端节点之间只有一条路由时，在一个复杂的网络拓扑结构中找到这一路由也不是一件容易的事。

网络管理员可以凭借经验（“人工地”）确定路由，他往往会基于各种不同的、没有被标准化的考虑。选择特定路由的理由可能是：特定类型的应用程序对网络的特殊需求、决定使用特定网络服务提供商的网络来传输信息、关于特定网络链路高峰负载的假设，以及基于安全的考虑。

然而，对于具有复杂拓扑结构的大型网络，根据经验确定路由的方法往往并不合适。在这样的网络中，人们使用自动化的方法确定路由。

为了达到这一目的，终端节点和其他网络设备都配备了特殊的软件工具，这些工具组织服务报文的交换，允许每个网络节点安排它自己对网络的描绘。然后，基于收集到的数据，使用不同的专门软件，自动确定合理的路由。

可以选择不同类型的网络信息来选定最优路由。然而，在解决这一问题时，往往只会考虑网络拓扑结构上的信息。图2-13描述了这一方法。在终端节点*A*和*C*之间，有两条路由可以传输信息：*A*-1-2-3-*C*和*A*-1-3-*C*。如果，除了节点之间的链路，不考虑其他关于网络的信息，那么选择*A*-1-3-*C*路由是一个合乎逻辑的选择。

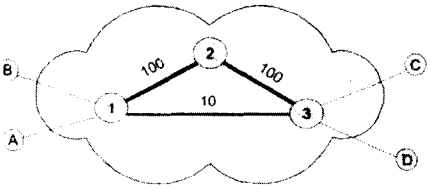


图2-13　路由选择

这一方法是基于最小化选择标准。

用来进行这一选择的路由参数称为**路由度量**（route metric）。

在这个例子中，最小化标准是路由的长度，以转发节点的数量来衡量。路由度量的最小化是选择路由的主要方法。

然而，很可能这一选择并不是非常合理。图2-13所示的设计显示，链路1-2和链路2-3的带宽为100Mb/s，而链路1-3的带宽为10Mb/s。因此，如果我们希望以最快的速度传递信息，选择*A*-1-2-3-*C*路由会更好，尽管它要经过三个转发节点。

出于这一原因，衡量路由长度可以选择不同的标准，包括转发节点的数量（正如刚才的例子）、路由的线性长度，甚至它的费用。为了建立例子中的标准，当信息必须以最快的速度传递时，每一条链路都有一个和它的带宽成反比的值。为了将它们作为整数操作，人们往往选择一些比网络链路带宽值更大的常数。例如，如果100Mb/s被选为这个常数，那么，链路1-2和链路2-3的标准被定为1，链路1-3的标准为10。路由的标准等于构成路由的链路的标准之和；因此，路由的1-2-3部分的标准为2，路由的1-3部分的标准为10。有最小标准值的路由是合理的路由，也就是，*A*-1-2-3-*C*。

这些选择路由的方法只是考虑了网络的拓扑结构，并没有考虑通信链路的负载[⊖]。以道路交通为例，比如说我们使用地图选择了一条行车路线，考虑了需要经过的城镇的数量，以及道路的宽度（类比于链路的带宽），我们在同等情况下更倾向于高速公路。然而，我们没有注意广播或是电视节目中关于当前交通阻塞的通知。因此，很可能我们的选择远非最优，尤其当许多流量已经在*A*-1-2-3-*C*路由上传输，而*A*-1-3-*C*路由是空闲的时候。

当路由选定后（人工或者自动），我们需要将这个选择通知所有网络设备。通知网络设备的报文必须将以下信息的变动通知每一个中间设备：“任何时候，网络设备接收到和数据流*N*相关的数据，都必须将它们传输到接口*F*，以便进一步转发。”网络设备处理每一个这种类型的路由报文。因此，新的记录被放入交换表，在交换表里，流的本地属性或全局属性（例如标记、输入接口号，

⊖ 使用通信链路当前负载信息的方法能找到更有效的路由。然而，它们也会造成网络节点更频繁地交换额外的辅助信息。

或者目的地址）被映射为设备需要转发这个流数据的接口号。

表2-1是交换表的一部分。它包含了指示节点将流M转发到接口G，流N到接口F，流P到接口H的记录。

表2-1 交换表的一部分

流属性	数据转发（接口号或下一节点的地址）
M	G
N	F
P	H

自然地，路由报文结构和交换表内容的详细描述取决于特定的网络拓扑结构。然而，这些特性并不改变过程的本质。

将路由信息（例如路由选择）传输到转发节点，可以人工或自动地完成。网络管理员可以通过人工配置设备改变特定的路由——例如，在很长一段时间内物理连接一组特定的输入输出接口。这和电话接线员操作第一代交换机类似。除此之外，网络管理员可以通过将需要的记录输入交换表来手工修改交换表。

然而，网络拓扑结构和信息流不时地在变化。这些变化可能由于节点失效或新的转发节点出现等因素而产生。此外，网络地址可能会变化，或者定义了新的流。因此，灵活地解决路由问题的方法，意味着连续地分析网络状态，并根据需要更新路由和交换表。在这些例子中，确定路由的任务必须由复杂的软件和硬件完成。

2.5.3 数据转发

当路由被确定并被记录在所有转发节点的交换表中后，所有的事情都已经准备完毕，等待执行最主要的操作——在终端节点之间进行实际的数据传输，或称为终端节点的交换。

对于每一对终端节点来说，这一操作可以表示为一系列本地（*local*）交换操作（它们的数量对应于转发节点的数量）的组合。也就是说，发送端必须将数据提供到特定的接口，所选择的路由从这个接口开始，所有转发节点必须相应地将数据从一个接口转发到另一个接口，换句话说，转发节点必须进行*本地接口交换*（*local interface switching*）。

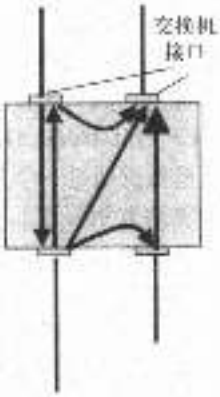


图2-14 交换机

进行交换操作的设备称为**交换机**（switch）（图2-14）。

然而，在进行交换操作之前，交换机必须辨别出这个流。为了达到这一目的，它必须分析传来的数据，发现交换表中特定流的属性。如果发现匹配，这些数据便被转发到路由中定义的接口。

要点 在电信网络中，诸如交换（switching）、交换表（switching table）、交换机（switch）等术语可以有多种解释。我们已经将交换定义为通过转发节点连接网络的过程。这一术语也被用来表示在一个转发节点中，接口之间的连接。广义上的交换机指任何可以在接口之间进行交换数据流操作的设备。交换操作可以根据不同的规则和算法完成。有些交换的方法以及它们相应的交换表和交换设备有特殊的名称。例如，对于网络层的技术，如IP和IPX，类似的概念有特殊的术语——路由（routing）、路由表（routing table），以及路由器（router）。对于以太网，交换表常常被称为转发表（forwarding table）。其他特殊类型的交换和它们相应的设备也被指定了同样的名称——狭义的交换、交换表和交换

机，例如，局域网交换机（LAN switch）和局域网交换（LAN switching）。对于电话网络，它们的出现远早于计算机网络，类似的术语也非常典型。这里，交换机和电话交换机同义。由于电话网络历史悠久、使用广泛，因此在电信领域交换机这一术语通常指电话交换机。

专用的设备和带有内置交换软件的通用计算机都可以担任交换机的角色。计算机可以将交换功能和正常的终端节点功能结合在一起。然而，在大多数情况下，使用专用的网络节点专门进行交换显然更为实际。这些节点构成了交换网络，所有其他的节点都连接在它们之上。图2-15显示了由节点1、5、6、8构成的交换网络，节点2、3、4、7、9、10、11连接在这个交换网络之上。

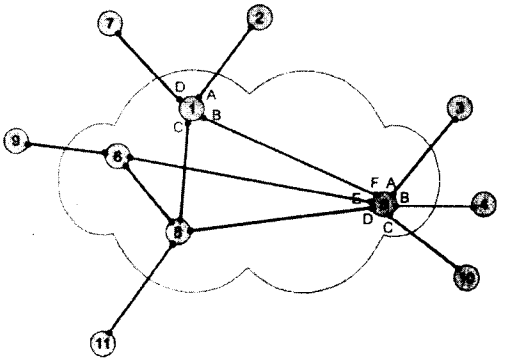


图2-15　交换网络

2.5.4　多路复用和解多路复用

为了判断到来的数据被转发到哪个接口上，交换机必须确定它们和哪个流相关。这一任务必须独立解决，无论发送到交换机输入端的是“单一”的流还是“混合”的流。“混合”的流是若干数据流聚集的结果。在这种情况下，流的识别必须先完成流的解复用，换句话说，将聚集的流分解为多个流。

通常，交换操作伴随着反向的多路复用操作，在这一过程中，多个数据流形成一个聚集的流。这一聚集的流可以使用单一的物理通信链路进行传输。

多路复用和解复用与交换同样重要。如果没有这些操作，就需要为每个流提供一条单独的链路。这就需要网络中有许多并行的链路，这将使部分连接网络的优势大打折扣。

图2-16描绘了三个交换机组成的一个网络的一部分。交换机1有五个网络接口。我们来考虑int.1接口上发生的事。它从三个接口——int.3、int.4和int.5接收数据。我们需要将所有这些数据通过一个共同的链路传输（也就是进行多路复用操作）。对于网络终端节点之间的多个网络会话，多路复用保证了已有物理链路的可用性。

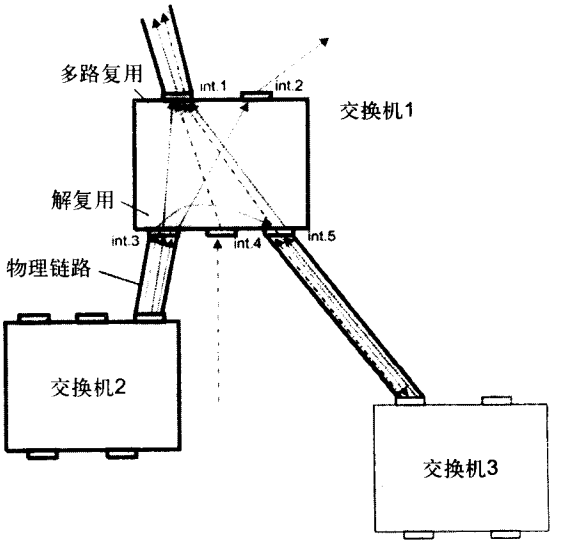


图2-16　流交换时的多路复用和解复用操作

在单个物理链路中，有多种流的多路复用方法，其中最重要的是**时分多路复用**（Time Division Multiplexing, TDM）和**频分多路复用**（Frequency Division Multiplexing, FDM）。在使用时分多路复用时，每一个流在一个固定的或随机的时间间隔内使用链路传输数据。当使用频分多路复用时，每一个流在分配给它的频率范围内传输数据。

多路复用技术必须允许这个聚集流的接收端进行一个反向操作——将数据解复用为单独的流。例如，在int.3接口上，交换机将聚集流解复用为三个单独的流。第一个被转发到int.1接口，第二个被发送到int.2接口，第三个被发送到int.5接口。对于int.2接口，由于这一接口只供一个单独的流使用，因此没有必要进行多路复用和解复用操作。事实上，多路复用和解复用可以在每一个支持双

工模式的接口上同时执行。

当所有到来的信息流被交换到一个单独的输出接口时，在那里，它们被多路复用为一个单个的聚集流，并且被转发到公共的链路上，这个交换机称为多路复用器。图2-17a显示了这种类型的交换机。具有单个输入接口和多个输出接口的交换机称为解复用器。(图2-17b)。

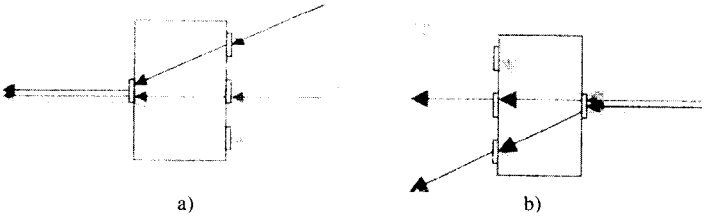


图2-17　多路复用器和解复用器

2.5.5　共享介质

连接到一条物理链路上的网络节点的数量是链路的另一个参数。在上面的例子中，只有两个交互的节点（准确地说，两个接口）被连接到通信链路上（图2-18a和图2-18b）。在电信网络中，使用了另一种类型的连接，其中，多个接口连接到一个单一的链路上（图2-18c）。这样的多个接口的多重连接导致了之前介绍过的公共总线技术，有时候称为菊花链连接（*daisy-chain connection*）。在所有这些情况下，人们必须解决多个接口间链路共享的问题。

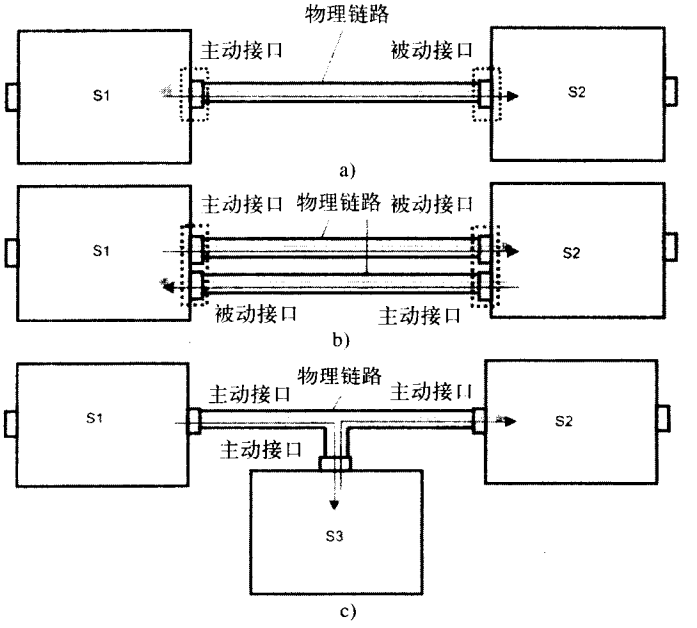


图2-18　通信链路的共享使用

图2-18显示了多个接口间共享链路的几种方式。在图2-18a中，两条单向物理链路（即每一条链路只能够以一个方向传输信息）连接了交换机S1和交换机S2。在这个例子中，传输接口是主动的，介质完全受这个接口的控制。被动接口仅仅接收数据。此时，两个接口之间共享链路没有任何问题。然而，请注意，在这个链路中，仍然需要解决数据多路复用的问题。事实上，如果两个设备间的两条单向链路实现了全双工的功能，那么它们被认为是一条双向链路。同样，单个设备的两个接口被认为是一个接口的发送和接收部分。

在图2-18b中，一条可以双向传输数据的链路连接了交换机S1和交换机S2，只是要轮流传输。我们需要实现S1、S2接口到这条链路的同步访问机制。图2-18c显示了其配置，其中，多于两个的接口连接到这一通信链路上，构成了一个公共总线，这是这个例子的一个推广。

用来供多个接口同时使用的物理链路称为**共享链路**（shared link）[⊖]。通常，我们也使用另一个术语：**共享介质**（shared medium）。共享通信链路不但用作交换机和交换机之间的链路，也用作计算机和交换机之间、计算机和计算机之间的链路。

有几种方法可以解决共享通信链路的多路访问问题。其中的多种方法使用了集中式的方法，由一个称为**仲裁器**（arbitrator）的特殊设备控制访问；其他的方法基于非集中式的方法。在不同模块间共享连接线路的问题也存在于单个计算机中。受处理器或特殊总线仲裁器控制的系统总线访问便是一个很好的例子。由于信号的传输需要更长的时间，组织网络内通信链路的共享访问有一些特殊的需求。出于这一原因，协调通信链路的访问过程需要更长的时间间隔，并导致网络性能的显著损失。因此，在广域网中，几乎从不使用介质共享。

在局域网中，由于实现简单和高效，介质共享更为常用。这一方法在以太网这一当前最流行的局域网技术中被广泛使用。同时，它也在过去曾经非常流行的令牌环和光纤分布式数据接口（FDDI）技术中使用。

然而，在最近几年中，出现了另一个趋势——放弃共享介质，即使是在局域网中。这一方法的最大优势——低成本，也同样导致了网络性能的损失。

要点　具有大量节点的共享介质网络总是比相类似的点对点连接网络慢，因为共享通信链路的带宽被多个网络中的计算机平分。

然而，通信线路的访问共享不但被经典网络技术保存下来，也被一些新的局域网技术所应用。例如，千兆以太网（在1998年作为一个新的标准被采纳）的开发者们将介质共享模式纳入其规范，同时，规范中也包括单独的连接链路。

2.5.6　交换类型

对通用交换问题的复杂技术解决方案构成了网络技术的基础。总而言之，对**每一个**特定交换任务的解决方案取决于这一集合中**其他**任务所选择的解决方案。交换任务包括：

- 确定流和合适的路由
- 构造交换表
- 辨别流
- 在同一设备的不同接口间传输数据
- 流的多路复用和解复用
- 介质共享

在解决交换问题的所有可能的方法中，应该区分以下两个基本方法：

- 电路交换（*circuit switching*）
- 分组交换（*packet switching*）

由于电路交换网络起源于第一代的电话网络，所以它们具有很长的历史。分组交换网络相对来说比较年轻。随着第一代广域网实验的开展，分组交换网络在20世纪60年代晚期出现。每个网络都有其自身的优点和缺点，但是，根据专家们对未来的预测，由于分组交换更通用更灵活，因此未来属于分组交换技术。

⊖ 我们需要强调的是，术语“共享介质”在传统意义上与多个接口之间的链路共享有关，它实际上从未用来描述流之间的链路共享（即多路复用和解复用）。

示例　让我们使用邮政服务的例子来描述通用交换模型。

1. 邮政服务用流操作。此时，流由邮件所组成。我们规定，收件人地址作为流的主要属性。为了简化起见，考虑将目的地国家作为唯一的地址属性：印度、挪威、巴西、俄罗斯等。有时候，关于可靠性和投递速度的特殊要求作为附加的流属性。例如，如果一封投递到巴西的邮件标有“航空信件”，那么，必须从所有投递到巴西的邮件流中区分出一个用航空信件投递的子流。

2. 对于每个流，邮政服务必须定义一条路由，邮件将通过路由上的一系列邮局，邮局类似于网络的交换机。邮政服务的悠久历史使得它对大多数的目的地址有预定义的路由。新的路由也有可能会出现。这可能是由于新的交通工具的出现或政治、经济的变化，等等。在选定了一条新的路由后，需要将这一情况通知整个邮局网络。显然，这些动作和电信网络的操作非常相似。

3. 邮件传递所选择的路由信息在每个邮局中都会被张贴出来，它们以表格的形式显示了在递送过程中，目的国家和下一个邮局的映射关系。例如，在布鲁塞尔的中央邮局，所有要递送到印度的邮件都会先被转递到罗马的邮局；要被递送到东京的邮件会先被递送到莫斯科中央邮局。这种邮政路由表和电信网络的交换表类似。

4. 每一个邮局的操作与交换操作类似。所有来自客户和其他邮局的邮件都先被分类，这表明进行了流识别。之后，属于同一流的邮件被打包到一起，根据交换表，可以找到下一个邮局。

小结

- 为了使网络用户能够访问其他计算机的资源，例如磁盘、打印机、测绘仪，我们需要使所有网络计算机配备特殊的工具。每一台计算机之所以能够将数据传输到通信链路上，必须有特殊的硬件配合完成——网络接口卡（Network Interface Card, NIC），以及网络接口卡驱动程序这一控制它的软件模块。高层次的任务诸如产生对资源的请求，会由操作系统的客户机模块执行，完成请求则由操作系统的服务器模块执行。
- 即使在只有两台计算机的最简单的网络中，使用通信链路对数据进行物理传输还是有一些问题，例如编码和调制、发送和接收设备的同步、传输数据的错误控制等。
- 物理信道传输数据的重要参数包括网络负载、信息速率或吞吐量、容量和带宽。
- 在网络中连接多于两台的计算机时，必须解决选择拓扑结构的问题。这些拓扑结构是全连接型、星型、环型、公共总线型、层级树型以及混合型。编址方式可以是扁平的或层级的，数值的或者是字符的。同时，你必须选择交换机制，以及通信链路的共享机制。
- 在部分连接的网络中，用户之间的连接通过交换建立起来（即通过转发节点网络连接起来）。这时，我们需要解决如下问题：数据流和路由定义、每一个转发节点的数据转发、流的多路复用和解复用。
- 在交换的不同方式中，我们必须区分以下两种基本方法：电路交换和分组交换。

复习题

1. 什么信息在连接计算机接口和外部设备接口的链路上传输？
2. 设备接口包括那些部分？
3. 在和外部设备交换数据的时候，操作系统会执行哪些操作？
4. 外部设备的驱动程序通常会做哪些操作？
5. 定义拓扑结构。

6. 若干节点连接成一个三角形，可以归类为哪种拓扑结构？
7. 若干节点连接成一个正方形，可以归类为哪种拓扑结构？
8. 三个顺序连接的节点（最后一个并没有连接到第一个）构成的结构，可以归类为哪种拓扑结构？
9. 公共总线型拓扑结构是以下哪一个的特例？
   A. 全连接型
   B. 环型
   C. 星型
10. 哪种拓扑结构具有增加可靠性的特点？
11. 哪种拓扑结构在今天的局域网中使用得最为广泛？
12. 编址系统的要求是什么？
13. 以下地址可以归类于哪种类型？
    www.olifer.net
    20-34-a2-00-c2-27
    128.145.23.170
14. flow和stream这两种流的区别是什么？
15. 哪些属性可以用来作为流的特征？
16. 描述路由选择时的主要方法和标准。
17. 以下哪些陈述在某些情况下是成立的？
    A. 通过连接一对接口，交换机固定了路由。
    B. 路由由网络管理员确定，并手工输入特殊的表格。
    C. 路由表在生产工厂里就被输入到交换机里。
    D. 路由表由网络硬件和软件自动生成。
    E. 每个交换机都在其上存储着一个特殊的路由表。
18. 这些设备中哪些可以被称为交换机？——自动电话交换机、路由器、网桥、多路复用器。
19. 多路复用使用哪些方法？
20. 解释介质共享和多路复用的区别。

练习题

1. 在任何类型的通信网络中，为了保证两个用户之间的信息交换，需要解决的主要问题是什么？
2. 请解释为什么将公共交通分成几个不同的流可以优化城市交通系统的控制。
3. 假设在网络终端节点A和B之间有几条路由。考虑在终端节点间不同数据传输的优点和缺点。
   - 使用所有已有路由并行传输数据，是否要优于根据一个特定标准优化的单个路由？
   - 使用几种可能的路由，并且在它们之间共享数据传输。哪些规则可以被用来定义转发下一个分组所需要的路由？

第3章 分组和电路交换

3.1 引言

在这一章中，我们将继续研究电信网络中的一般交换原理。我们将首先关注两种最主要的通信原理并和它们进行比较——电路交换和分组交换。

电路交换的出现远早于分组交换。它的工作原理起源于第一代电话网络。电路交换原理的主要局限是，无法动态重新分配物理链路的带宽。

分组交换原理由计算机网络的工程师发明。它考虑了计算机数据传输的特征。例如，数据的突发性。比起传统的电话网络中的电路交换方法，它更适合于计算机网络。

然而，网络技术中任何优点和缺点都是相对的。在传输突发性数据时，使用分组交换网络交换机中的缓存，可以更为有效地利用链路的带宽。但这也导致了分组传递中不确定的延迟。对实时数据来说，这种延迟是一种不足。这类数据习惯上都使用电路交换技术来传输。

本章介绍了分组交换网络中二种最常用的分组转发方法：数据报传输、面向连接的传输和虚电路技术。

本章的最后，我们将介绍局域网中广泛使用的介质共享原理。

3.2 电路交换

首先，我们考虑简化的电路交换，这将有助于解释这一方法的基本思想。如图3-1所示，交换机用通信链路连接，构成了交换机网络。每一条链路都有相同的带宽。

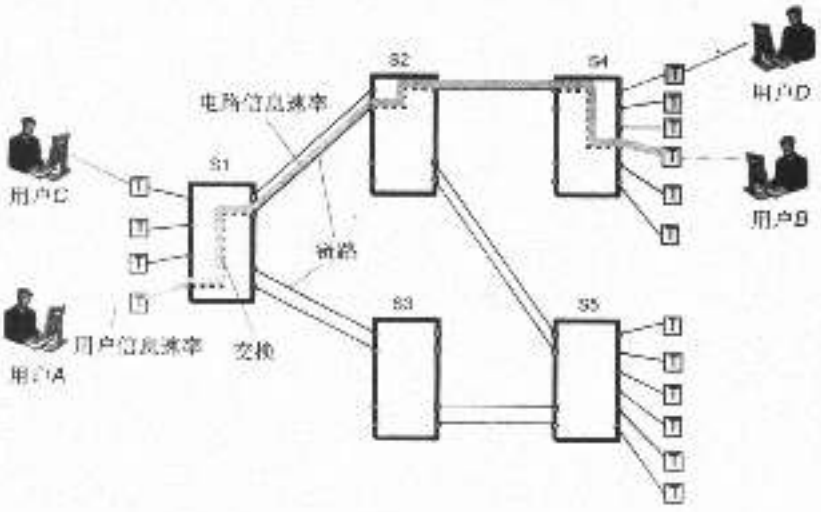


图3-1 不带多路复用的电路交换

每一个终端节点（用户）使用终端设备连接到网络，这些终端设备以等于链路带宽的恒定的速度，向网络发送信息。如果在某段时间内，负载比链路带宽低，终端设备通过在有用的用户信息中加入空白的无意义的数据的方式，以一个恒定的速率持续向网络发出数据（图3-2）。终端设备

知道位流的哪部分是有用信息，哪部分是填充信息。接收端的终端设备必须丢弃无意义的信息，只将发送端发送给网络的数据提供给用户。

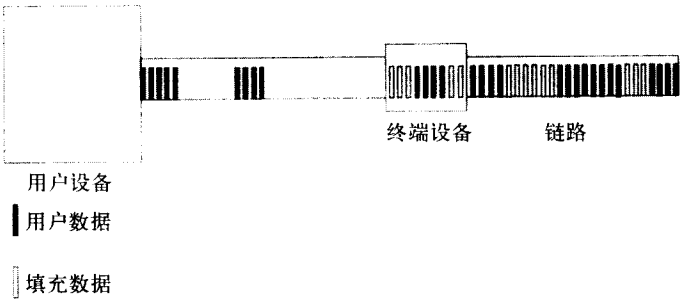


图3-2　将流补足到链路带宽

由于大多数人已经习惯了电话网络这一电路交换网络最著名的代表，我们的介绍将引用一些电话技术中的功能特征。

3.2.1　连接建立

数据交换在**连接建立**（connection setup）之后开始。

假设两个电话用户（*A*和*B*）希望交换一些数据。在向网络发送数据前（即，开始对话），用户*A*向交换网络发出一个请求。在这个请求中，需要指定用户*B*的地址（即，电话号码）。发送这一请求的目的是在用户*A*和用户*B*之间通过一条信息链路建立连接，这一连接的性质类似于一条持续的通信链路。它以一个固定的速度，在整条链路长度上传输数据。这意味着中间交换机不需要缓存用户数据。

为了建立这样一条链路，请求必须通过从*A*到*B*的一系列交换机，保证所有要求的路由部分（通信链路）都可用。除此之外，为了成功建立这一连接，终端节点*B*必须是空闲的（即，并不忙于另一个已建立的连接）。为了固定这一连接，在从*A*到*B*的路径上，每一个参与的交换机都存储着一些信息，为*A*-*B*之间的连接保留着适当的路由信息。在每个交换机内部，建立起了与数据路由相对应的内部接口之间的连接。

3.2.2　建立请求的阻塞

建立请求的阻塞（setup request blocking）是电路交换技术的一个重要特征。如果*A*和*B*之间已经建立了一个连接，这时，任何其他用户请求网络建立一个连接，如果这一连接需要*A*和*B*之间已经保留的连接路由的至少一个部分，那么，网络就会拒绝这一请求。例如，如果终端节点*C*发送一个请求，希望建立一条到终端节点*D*的连接，网络会阻塞这一建立请求，因为连接交换机S2和交换机S4之间的唯一链路已经被保留给用户*A*和用户*B*。

连接建立的阻塞也可以发生在路由的终端部分。例如，如果被呼叫的用户已经连接着另一个终端节点时，便会发生这种情况。当发生这种情况时，网络会通知被呼叫的用户，告诉它这一事件。这和电话网络类似，电话网络用连续的短音答复（或者是**线路正忙**（line busy）的信号）。有些电话网络可以分辨出不同的事件，例如**网络忙**（network busy）或者**用户忙**（subscriber busy），然后通过使用不同频率或不同声调的短音，将这一事件通知呼叫用户。

3.2.3　保证带宽

假设用户*A*和用户*B*之间的连接已经建立。现在，只有这两个用户可以使用这一固定带宽的电路。这意味着，在整段连接时间内，它们必须以固定速率向网络发送数据；同时，网络保证以同

样的速率无损地将这些数据传输到被呼叫的用户。这与此时网络中是否存在其他连接无关。用户不能以超过线路带宽的速率向网络传输数据。网络也不能降低用户数据的传输速率。

网络负载只会影响建立请求阻塞的可能性。网络中建立了越多的连接，越有可能发生建立请求阻塞。

网络以一个小的、固定的延迟进行数据传输其实是一件好事。电路交换网络的小的、固定的传输延迟，保证了对高延迟非常敏感的数据的高质量传输。这也被称为**实时流量**（real-time traffic），其典型代表是语音和视频。

3.2.4 多路复用

我们已经介绍了简化的网络，其中，每一条物理链路总是以恒定的速率传输数据，该简化网络工作起来并不高效。

首先，这类网络的用户通常不能得到它们所要求的服务。它们必须是普通的、标准的用户，只能以可以获得的固定速率传输信息。今天，很难想像只有这样的用户，尤其是，当不同类型的终端设备得到广泛使用时，例如固定和移动的电话和计算机。因此，总体而言，用户流量的速率与物理链路的固定带宽并不相同。带宽可以显著地超过或者低于用户的需求。对于前者，用户不能最大限度地利用电路；对于后者，用户只能或限制需求，或使用多条物理链路。

其次，网络不能有效地使用它的资源。显然，图3-1中的网络在交换机之间没有足够数量的链路。为了将阻塞减小到可以接受的程度，我们需要在交换机之间放置大量并行的线路。这是一种昂贵的方法。

为了提高电路交换网络的效率，人们使用了多路复用，多路复用使用一条物理链路，从多个逻辑连接同时传输数据。电路交换网络中的多路复用有特定的特点。例如，每条链路的带宽被划分为**相同部分**（equal part），这样，就提供了相同数量的所谓**子信道**（subchannel）。请注意，为了简化起见，子信道也常常被简单地称为信道。通常，比起连接交换机之间的链路，用户和网络之间的通信链路支持较少数量的信道。在这种情况下，阻塞的可能性将被减少。例如，用户链路可以由2、24或30条信道组成；交换机之间的链路可以由480或1 920条信道组成。目前，最常见的数字子信道的速率是64Kb/s。这一速率可以保证以数字形式传输的语音的质量。

当电路交换网络增加了多路复用机制之后，它的工作机制就改变了。请求建立逻辑连接的用户现在只保留链路的一个或者多个子信道，而不是整条链路。这样，连接被建立在子信道层，而不是在链路层。可能会保留多个子信道，以防单个信道的带宽不够。这使得用户可以保留与所需数据传输速率最接近的子信道（或多个子信道）。此外，多路复用使得在交换机之间可以建立更有效的链路。为了减小阻塞的可能，可以使用具有大量逻辑子信道的单一物理链路，而不是几条物理链路。

图3-3是带有多路复用的电路交换网络。网络中建立了两条链路，*A-B*和*C-D*，前者在每一条通信链路中使用一个子信道，后者在每个链路中使用两个子信道。因此，虽然有图3-1的网络结构，*C-D*之间的第二个呼叫并不会被阻塞，因为交换机支持多路复用。

> **注意** 在使用多路复用时，电路交换网络的基本性质是不变的，即之前使用的由多个路由部分带宽构成的**聚合信道**（aggregate channel），或叫电路。现在，唯一的区别是子信道担任了链路的角色。

显然，使用多路复用使交换机端流量的处理过程更加复杂。现在，需要将数据传输到需要的信道，而不是在端口间简单直接地交换。在使用时分多路复用时，需要在两个信息流之间做一个高层次的同步。在使用频分多路复用时，需要使用频率转换。

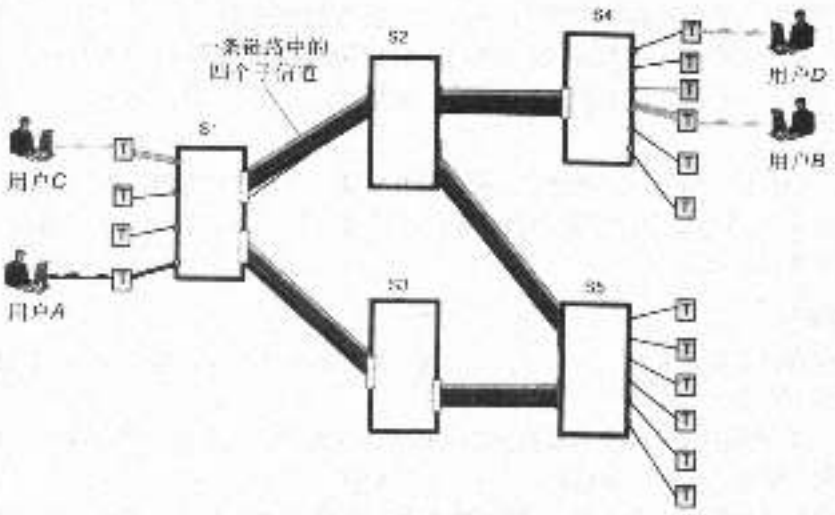


图3-3　带有多路复用的电路交换

3.2.5　传送突发流量的低效率

电路交换网络的低效率还有另外一个原因，这源于这类网络的基本运行原理，也就是，在整个连接时间内，保留电路的固定带宽。

我们已经介绍过，多路复用提高了电路交换网络的效率，因为用户可以根据需要选择连接速率。然而，这只对那些以固定速率产生信息流的用户有效。假如用户的信息流是突发性的呢（即在到网中到达）？例如，向网络发送数据，然后是一段空闲时间。

如果你仔细考虑用户流量，你会发现，事实上所有电信网络的用户都属于这一范畴。让我们回忆一下，电话网络的用户以恒定的速率传输信息。这一表面上的恒定是这样达到的，电话网络的终端设备处理不平均的用户数据流。例如，数字电话以64Kb/s这一固定速率传输信息，无论用户是否在说话。显然，如果电话可以把对话中的停顿消除出去，只把有用的信息传送给网络，那它会工作得更加有效。

最后，还有另一类的用户，非常明显，他们对信息传输的需求是可变化的速率。他们就是计算机用户。

浏览互联网的用户行为产生了突发流量。在下载网页到用户计算机的过程中，流量速率增长迅速，当下载过程完成后，流量速率下降到几乎为零。这一过程不停地重复。

个人网络用户的流量脉冲系数（traffic pulsation coefficient）等于数据交换的平均强度和最大可能强度的比值。这一系数可以达到1：100。如果用户计算机和服务器之间的电路交换建立起来之后，在这一会话的大部分时间里，电路是无效运行的。另一方面，网络性能的某些部分会专属于这对终端节点。因此，其他网络用户就无法获得它们。这段时间的网络运行，可以比作地铁站中无人乘坐的自动扶梯。电梯始终在运作，虽然它并没有完成任何有用的工作。

当流量在整个会话过程中具有恒定的强度，并且与网络物理信道带宽一致时，电路交换网络能最有效地传输用户信息。

任何网络技术的优势和劣势都是相对的。在有些情况下，优势显示出来，而劣势变得微不足道。因此，在仅仅需要传输电话流量的情况下，电路交换技术非常有效，因为无法从对话中消除（cutting off）停顿是可以被容忍的。然而，在传输计算机数据时，由于这些数据有突发性的性质，电路交换的低效率就不再是无关紧要的了。

3.3 分组交换

分组交换（packet switching）技术是专门为计算机数据的有效传输而设计的。

在使用分组交换时，网络用户的所有传输数据都被分为相对较小的片段，称为**分组**（packet），有时候也称为**帧**（frame）或**信元**（cell），虽然在这种情况下，术语的选择无关紧要。这一划分分组的操作在传输节点完成（图3-4）。每一分组都有一个头部（header），头部包含地址，我们需要通过这一地址将分组传送到目的节点。每个分组中包含地址是分组交换技术的基本性质之一，因为每个分组才可能[⊖]独立于信息流中的其他任何分组，被某个交换机处理。除了头部之外，分组还有另外一个辅助的域，它往往在分组的末尾，因此常被称为**尾部**（trailer）。尾部包含校验和，它可以让校检信息在传输的过程中是否被破坏。

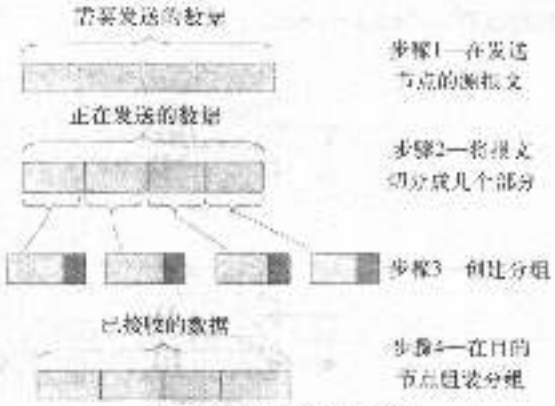


图3-4 数据流划分为分组

分组发送到网络时，并不事先保留通信链路，分组以数据源生成数据的速率传输到网络。这一速率不能超过接入链路的带宽。与电路交换网络相反，分组交换网络被认为能随时从任何终端节点接收数据。

注意 带宽保留这一概念也能在分组交换网络中得到应用，但是，这种保留的主要思想与电路交换网络中的带宽保留不同。区别在于，根据每一信息流的当前需求，分组交换网络的信道带宽可以在信息流间动态地重新分配，电路交换网络没有这种可能性。第7章将更详细地介绍这一带宽保留技术的细节。

3.3.1 缓存与队列

与电路交换网络一样，分组交换网络由被物理数据链路连接起来的交换机组成。但是，在这两种网络中，交换机的工作方式并不相同。

其中，最主要的区别是，分组交换机带有内置缓存，这些缓存可以暂时存储分组。

首先，这是由于交换机需要分析的所有部分，才能做出转发的决定。这些部分包括带有目的地址的头部、数据域、以及带有校验和的尾部。交换机检查校验和，并且，只有当分组数据没有被破坏时，它才开始分组处理。也就是，交换机根据目的地址确定下一个交换机。因此，每一个分组都被放置到**输入缓存**（input buffer）中（也就是说，它被一比特一比特、顺序地放置到分配给这个分组的内存中）。如果我们考虑这一情形，我们可以说分组交换网络使用了**存储转发技术**（store-and-forward technique）。需要注意的是，为了实现这一目的，我们需要有一个和单个分组大小相同的缓存。

其次，缓存被用来协调报文到达的速度和它们交换的速度。如果进行分组交换的设备（交换结构）跟不上分组处理的速度，那么交换机的接口部分就会创建输入队列。为了存储这一输入队列，缓存的大小必须超出单个报文的大小。创建交换结构有不同的方法。传统的方法基于单个中央处理器，这一处理器为交换机的所有输入队列服务。这一方法可能会产生很长的队列，因为多个队列共享处理器的性能。当代建立交换结构的方法是使用多处理器，每一个接口都有它自己的

⊖ 回顾“有可能”，这里有其特殊的重要性。因为在不同的分组交换技术中，分组处理的完全独立性并不能被保证（例如，参见帧中继技术）。

内置处理器，用以进行分组处理。除此之外，还有一个中央处理器，它负责协调接口处理器的运作。使用接口处理器提高了交换机的性能，减少了输入接口队列的长度。但是，这样的队列仍然可能出现，因为中央处理器可能成为瓶颈，就像它以前发生的那样。第15章将详细介绍交换机内部结构的不同方面。

最后，缓存被用来协调连接到特定分组交换机的链路的速率。如果从某链路到交换机的分组速率超出了这些分组需要被转发的链路的带宽，那么，为了避免分组丢失，需要在目的接口建立**输出队列**（output queue）（图3-5）。

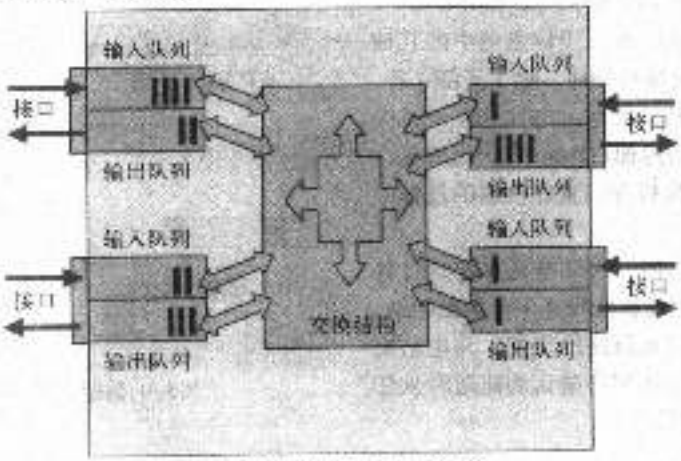


图3-5　分组交换机中的队列

于是，分组会在交换机的缓存中保存一段时间，然后，通过输出接口，它被转发到下一个交换机。这种数据传输的方式在交换机之间的主干链路上平缓了流量的突发性，这使得信道可以以最有效率的方式使用，增加了网络的整体性能（图3-6）。

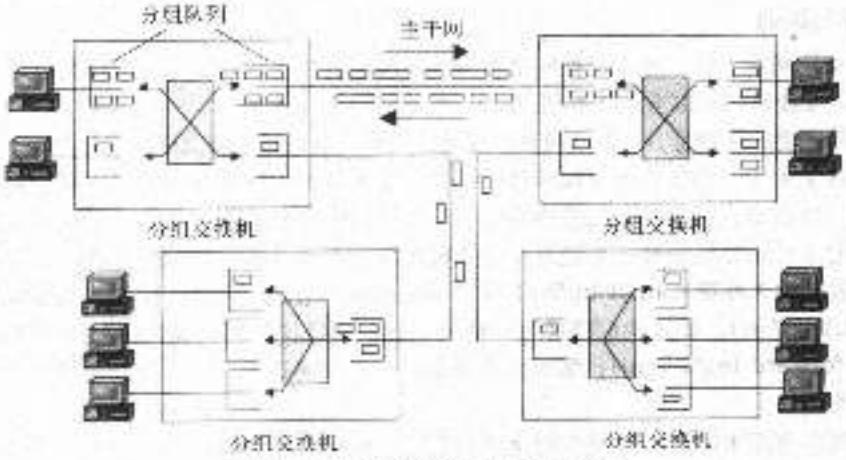


图3-6　平缓分组交换网络中流量峰值

根据大数定理，在一段时间段内，分组交换网络中众多个人用户突发性流量的峰值往往不会重合在一起。因此，当用户的数量足够多时，交换机的负载相当平衡。图3-6显示，从每一个终端节点到达交换机的数据在时间分布上非常不均。然而，更高层级的交换机（为低层级交换机之间的连接提供服务的交换机）负载更平衡。此外，连接更高层级交换机的主干链路的利用率接近上限。缓存技术使突发性流量更平缓。因此，主干链路的脉冲系数比用户接入链路低得多。

由于交换机中的缓存大小是有限的，有时候分组会丢失。**拥塞**（congestion）指某些网络部分的暂时超负载。通常，这发生在多个信息流的高峰恰好同时到达时。由于分组丢失是分组交换网络的内在性质，于是，人们开发了一套弥补这一不足的机制，用以保证这类网络的正常运作。减少丢包事件概率的方法是当前开发的热点。它们被称为**服务质量**（Quality of Service，QoS）和**流量工程**（Traffic Engineering），我们将在第7章中介绍这些技术。

3.3.2 分组转发方法

选择到达分组将要被转发的接口时，会基于以下三种**分组转发方法**（packet-forwarding method）：

- **无连接**（connectless），也被称为数据报传输。在这种情况下，传输的时候并不建立连接，所有需要被传输的分组使用同样的规则，一个接一个独立地转发。分组处理的过程只由分组内包含的参数值和当前网络状态决定。例如，根据当前的负载，分组可能在一个队列中停留更长或更短的时间。但是，网络不存储任何已经被传输的分组的信息，而且，在处理下一个分组时，这些信息也不会被考虑。这意味着网络把每一个分组当成数据传输过程中的独立的单元，称为**数据报**（datagram）。
- **面向连接的传输**（Connection-oriented transmission）。在这种情况下，面向连接的数据传输过程被划分为所谓**会话**（session），或叫做逻辑连接。网络对每一个逻辑连接的开始和结束进行记录。现在，作为会话一部分的整个传输分组，而不是每一个单个的分组，确定处理。处理每一个分组的过程直接取决于会话之前的历史。例如，如果之前多个分组被丢失，那么，传输后续分组的速率会被降低。
- **虚电路**（virtual circuit）。如果连接参数的列表包含路由，那么，所有需要在一个特定连接中被传输的分组必须使用一条特定的路由。这一单一的、固定的、事先被确定的、连接分组交换网络终端节点的路由，被称为虚电路，或者**虚信道**（virtual channel）。

图3-7对现有交换方法的分类进行了概括。

相同的网络技术可以使用不同的数据传输方法。例如，IP数据报协议用来在构成整个因特网的不同网络之间传输数据。因特网终端节点之间可靠的数据传输由面向连接的TCP协议完成，它将建立不含固定路由的逻辑连接。最后，因特网是使用虚电路网络的一个典型例子，因为它包含了许多ATM网络和帧中继网络，它们都支持虚电路。

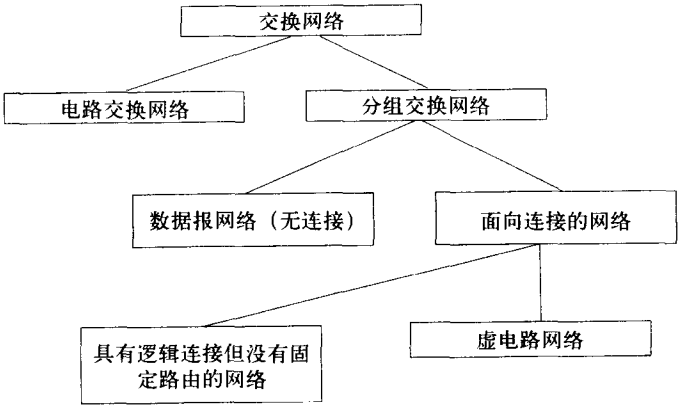


图3-7 分组交换网络的分类

3.3.3 数据报传输

正如我们之前介绍的那样，数据报传输方法基于这样一个事实：所有传输的数据报都独立地

被处理，在转发到达的分组时，对接口的选择建立在分组头部记录的目的地址（destination address），并不考虑一个特定的分组是否属于一个特定的信息流。

分组转发的方案基于**交换表**（switching table），它含有目标地址集和地址信息，这些信息唯一地决定了路由上的下一个网络节点（或是中间节点，或是终端节点）。让我们回忆一下，在不同的网络技术中，交换表有时被用其他的术语来指代。这些术语可以是**路由表**（routing table）、**转发表**（forwarding table），等等。为了更简单一些，从现在开始，我们将使用交换表这一术语指代这类表，这些表被用来进行基于目的节点地址的数据报传输。

数据报网络的交换表必须包含那些到达交换机接口分组可能会被转发的所有地址项。一般而言，到达的分组可能会被要求发送到任何网络节点。实际上，一些方法可以减少交换表中的项目数。其中一个便是层级地址，根据这一层级地址，交换表只需要包含地址的最主要（左边）部分，这些信息对应于一组节点（子网），而不是单个的节点。因此，我们可以类比于信件地址，对于信件地址而言，国家名和城市名对应于地址的最重要部分，很自然的，国家名和城市名在数量上要远远小于街道名、家庭地址，以及收件人的姓名，它们不是一个数量级的。

尽管使用了层级地址，在某些大规模网络中（例如，因特网），交换机中的交换表仍然可能有成千上万项。图3-8显示了数据报网络中交换机S1的交换表。

对于同一个目的地址，交换表可能含有多项，它们表明了下一个交换机的不同的地址。这一方法被称为**负载均衡**（load balancing），它被用于提高网络的性能和可靠性。在图3-8的例子中，分组到达S1交换机，最终要到达N2这一目的节点，它的下一交换机在S2和S3之间分配，这样可以平衡负载。这一做法减少了S2交换机和S3交换机的工作负荷，因此，在加速传输时，缩短了队列。具有相同目的地址的分组的路由有些**模糊**（fuzziness），这是独立分组处理原理的直接后果。这一现象是数据报方法所固有的。由于网络状态的改变，例如某些中间节点的失效，试图到达同一目的地址的分组可以通过不同的路由来传送。

S1交换机的交换表

目的地址	下一交换机点地址
N1	没有必要转发分组
N2	S2
N3	S3
N4	S3
N5	S6
N6	S6

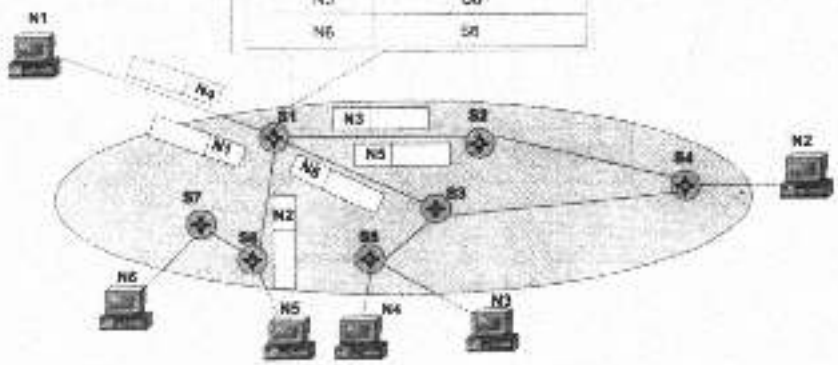


图3-8　分组转发的数据报原理

由于在数据实际传输之前，不需要做任何准备工作，数据报传输方式相对较快。但是，在使

用这种方法时，很难追踪数据报往目的节点的传输过程。因此，尽管网络尽最大努力传输数据报到目的节点，但这一方法并不保证分组一定能传输成功。类似这种类型的服务称为**尽力服务**(best effort service)。

3.3.4 逻辑连接

基于连接的传输会根据对之前的分组交换情况的了解而进行工作（例如，连接的一些当前值）。对于处理每一个新到来的分组，它提供了一种更合理的方法。连接参数可以有不同的用途。例如，分组编号和追踪发送、接收的分组数可以用来提高传输的可靠性。这将帮助丢弃重复的分组，对接收到的分组进行排序，在一个特定的连接中重传丢失的分组。一个安全连接的参数可能包含诸如加密方法等的信息。

连接的参数可以在整个连接时间内都是固定的（例如，最大分组大小），也可以为了动态反映当前连接状态而变化（例如，之前提到的分组的顺序号）。当发送端和接收端建立了一个新的连接时，它们首先协商交换过程的初始参数，之后才开始数据传输。

基于连接的协议提供更可靠的传输。但是，它们在数据传输时花费更多的时间，并且使终端节点有更高的计算负载（图3-9）。

在使用基于连接的传输时，源节点向目的节点发送一个具有特殊格式的服务分组，这一分组包含了建立连接的请求（图3-9b）。如果目的节点同意建立连接，它将回复给源节点另一个服务分组，确认连接建立，并建议一些参数用于这一逻辑连接。这些参数可能包括连接标识符、分组数据域长度的最大值，可以发送的、但未收到确认的最大分组数。发起连接的节点可以通过发送第三个服务分组来完成建立连接的过程，这一分组表明建议的参数可以被接受。之后，我们可以认为已经建立了一个逻辑连接。逻辑连接既可以用来进行单向数据传输（从连接的发起者开始），也可以用来进行双向的数据传输。在传输完某个完整的数据集合之后（例如，一个文件），发送节点通过发送一个服务分组，启动连接终止程序。

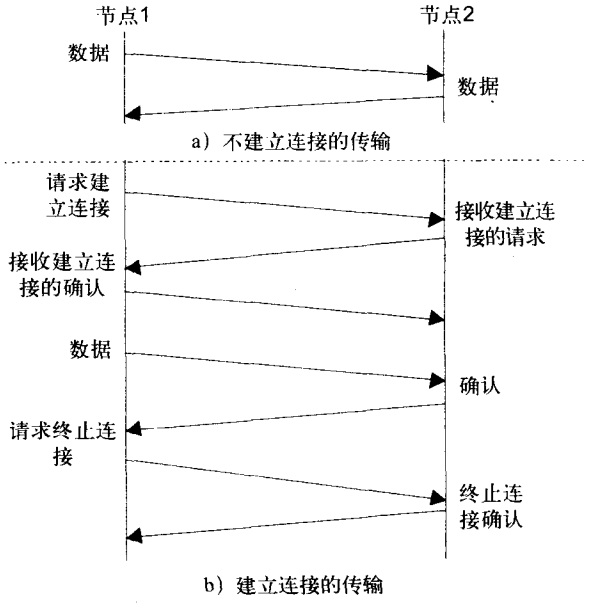


图3-9 传输

需要注意的是，数据报传输只支持一种分组——信息分组。和数据报传输相比，基于连接的传输至少需要支持两种类型的分组。它们是用来建立（或者中断）连接的服务分组，以及用来传输用户数据的信息分组。

3.3.5 虚电路

虚电路（virtual circuit）（**虚信道**（virtual channel））机制为分组交换网络中的数据传输创建固定的路由。与同一逻辑连接相关联的所有分组使用同一路由——虚电路。基于X.25、帧中继、异步传输模式的网络使用这一机制。

虚电路说明了网络中数据流的存在。为了在聚集流量中识别出一个数据流，这一流的每一个分组都必须进行特殊的标记。在这种类型的网络中，数据传输需要有一个前序过程——建立逻辑

连接。这一逻辑连接被称为虚电路。与建立逻辑连接的过程相似，虚电路的建立开始于源节点发出建立连接请求。建立连接的请求是一种特殊格式的服务分组，也被称为建立分组（set-up packet）。建立分组必须包含目的地址和虚电路建立后要传输的流的标识。建立分组通过网络传输，对于所有位于发送端到接收端路由上的交换机，建立分组会对控制信息进行登记。基于这些信息形成了路由表的表项。这些表项表明交换机该如何为带有这一标记的分组提供服务。用这种方式创建出的虚电路使用同一个标记进行标识[6]。

创建了虚电路之后，网络可以通过它来传输数据流。对于所有带有用户信息的分组，目的地址并没有被指定，信息分组仅仅包含虚电路标记，而不含目的地址。当分组到达交换机的输入接口时，交换机从到达分组的头部读取标记值，并查询交换表。然后，它将找出相应表项，这一表项表明分组将要被转发到哪一个输出端口。

虚电路网络中所使用的交换表与数据报网络中所使用的不同。在使用数据报转发算法的网络中，交换表包含所有可能的目的地址的信息。与之相反，虚电路网络中的交换表只包含通过这一交换机的那些虚电路的信息。通常，在一个大型网络中，通过一个特定节点的虚电路的数量要远远地小于节点的总数。这样，交换表的大小也会小得多。因此，查找所需要的表项耗时更少，也不需要占用交换机过多的处理能力。由于分组现在只需包含一个较短的数据流标记，而不是更长的目的地址，标记要比目的节点地址短得多。这样便减小了分组的额外信息长度。

> 注意　需要指出的是，在网络中使用虚电路技术并不表示网络成为了电路交换网络。虽然使用了建立电路这一步骤，但这一电路是逻辑的，它传输的是单个的分组，而不是像电路交换网络那样，以固定的速度传输信息流。

图3-10描述了网络的一部分，在这一网络中，建立了两条虚电路。第一条虚电路从具有N1地址的终端节点开始，到具有N2地址的节点结束，它通过了S1、S3、S4这些中间交换机。第二条虚电路保证了通过路由N3-S5-S7-S4-N2的数据转发。因此，在两个终端节点之间，可能有多条虚电路。

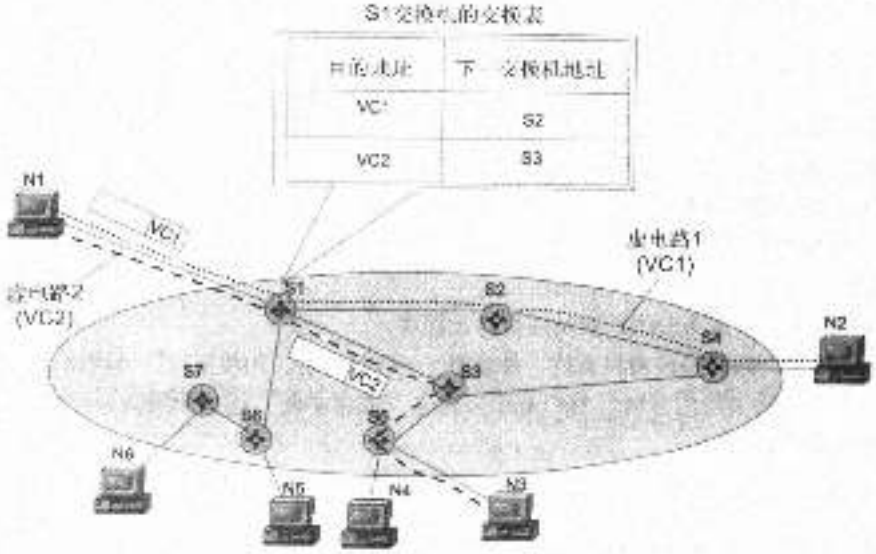


图3-10　虚电路的运行原理

[6] 在不同的技术中，这一标识有不同的名称：在X.25中，称为逻辑电路号（Logical Circuit Number，LCN），在帧中继中，称为数据链路连接标识符（Data Link Connection Identifier，DLCI），在异步传输模式中，称为虚电路标识符（Virtual Circuit Identifier，VCI）。

3.3.6 电路交换网络与分组交换网络的比较

在进行分组交换网络和电路交换网络的技术比较之前，我们先以交通为例，进行一个非正式的比较。

1. 电路交换网络与分组交换网络的运输类比

在使用这一类比时，车辆代表数据分组，道路和高速公路对应于通信链路。与数据分组相似，车辆独立地移动。它们共享道路空间，彼此会成为对方的障碍。如果交通过于拥挤和道路空间不相适应，则可能发生拥塞。于是，车辆在交通阻塞中耽误了时间，这一现象对应于交换机中分组的队列。

车流的交换发生于十字路口或是道路的交叉口，在那里，每一位司机选择一个合适的方向开往目的地。很自然地，和分组交换机相比，十字路口的角色是被动的，只有在用信号灯控制的十字路口，十字路口才能主动地参与交通控制，在这样的路口，信号灯为每一个穿越交叉路口的车流定义了转弯的方向。自然，如果由交通警察来完成这一任务，他的角色就更主动了，因为他可以从整个流中选出单个的车辆，并且允许其司机进行某项操作。

有关城市交通的类比可以用来比较分组交换网络和电路交换网络。

有时候，我们需要对某些车流保持特殊的状况。例如，假设有一个很长的车队要将孩子们带到夏令营营地。这一车队在高速公路上行驶，并且占据了多条车道。为了保证车队的行进没有障碍，我们需要事先选定路线。然后，在整条事先选定的、穿过若干个交叉路口的路线上，为这一车队指定一条单独的车道。交通警察为车队保留这一车道，保证这一车道不被其他车辆占用。只有当车队抵达了目的地点，这一保留才被取消。

在行驶的过程中，所有的车辆以相同的速度行驶，保持大致相同的间隔，以防止对其他车辆造成障碍。显然，对于这一车队，我们创建了一个具有特权的状况。但是，在这一例子中，车辆不再是单独地行驶，相反地，它们形成了流，其中的车辆不能向边上转向。在这一情况下，道路并没有得到有效的使用，因为车道在很长时间内没有被使用。这和电路交换网络中带宽的非有效使用相类似。

2. 延迟的定量比较

现在，让我们从传输的类比回到网络的流量。假设用户需要传输突发性的数据，这些数据由若干段的活动期和停止期构成。另外，假设用户可以选择使用电路交换网络或是分组交换网络传输。对这两种网络，通信链路的带宽是一样的。对这个用户而言，根据时间的要求，电路交换网络是最有效的。在电路交换网络中，用户有一个保留的通信电路，由他单独使用。使用这种方法时，所有数据都会无延迟地被递送到目的地。在大部分的连接时间内，被保留的通信电路将不能有效地使用（在没有数据的停止期间），但这对用户来说并不重要，因为用户的主要目的是越快越好地解决问题。

当用户决定使用分组交换网络时，数据传输的过程将会变慢，因为在传输的过程中，用户发送的到目的地的分组可能会在队列中延迟多次。对单个用户来说，分组交换网络会减慢传输速度，因为该用户的分组和其他用户发送的分组共享所有的网络资源。

让我们仔细考虑这两种网络数据传输延迟的原因。假设终端节点N1发送一个报文给终端节点N2。在数据传输的路由上，有两个交换机。

在电路交换网络中，由于建立电路造成一个起始延迟，之后，数据传输以电路的标准速率开始（图3-11）。将数据递送到目的节点所需要的时间（T）等于信号传播时间（t_{prg}）和报文传输时间（t_{trns}）之和。需要注意的是，交换机的存在对数据传输所需要的总时间没有影响。

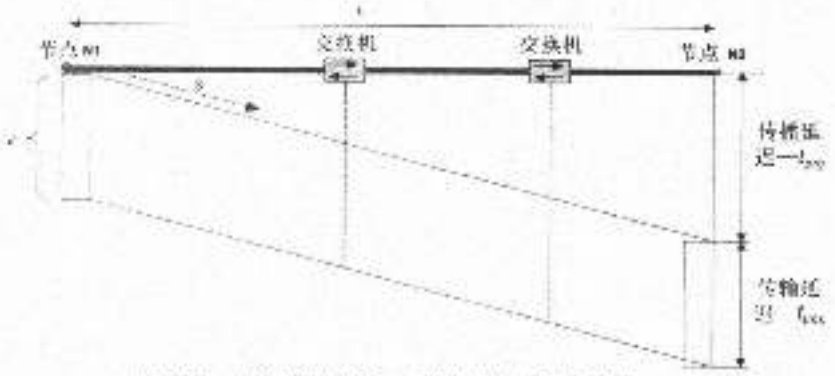


图3-11　电路交换网络中，报文传输的时间图解

注意　报文传输时间等于从信道中接收报文并放入目的节点缓存所需要的时间。在这一情况下，它被称做缓存时间（*buffering time*）。

- **信号传播**时间（signal propagation time）取决于源和目的的距离（*L*），以及电磁波在物理介质上的传播速度（*S*），它在$0.6v_{light}$到$0.9v_{light}$范围内变化，其中，v_{light}是真空中光的传播速度。因此，$t_{prop} = L/S$。
- **报文传输时间**（message transmission time）等于以比特为单位的报文大小（*V*）和以比特每秒为单位的电路带宽（*C*）的比值：$t_{trans} = V/C$。

在分组交换网络中，数据传输的过程不需要建立连接。

假设分组交换网络（图3-12）传输一个和上面例子（图3-11）中同样大小（*V*）的报文。但是，在这一情况下，报文被分成两个分组，每一个分组都带有一个头部。分组从N1节点传输到N2节点，这之间有两个交换机。每一交换机上的分组显示两次：当分组到达输入接口时显示一次，当分组从输出接口传输到网络时显示一次。很明显每一台交换机都会造成分组传输的一段时间的延迟。T_p代表在分组交换网络中，源发数据到目的节点所需要的时间，T_t代表在网络上传输单个（第一个）分组所需要的时间。

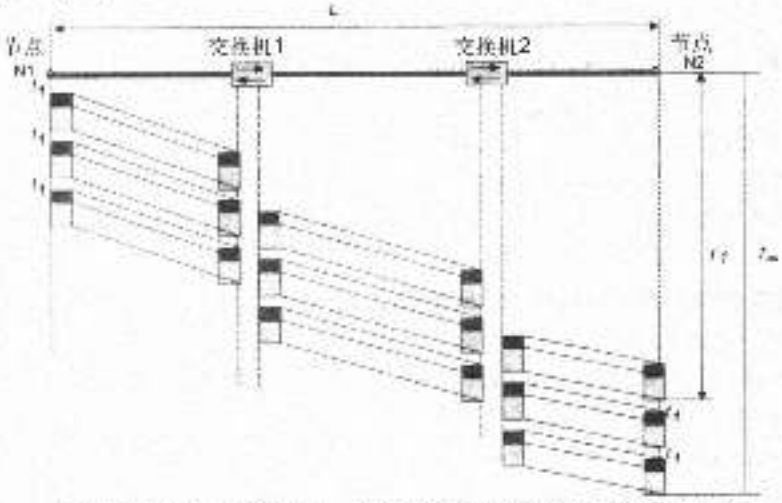


图3-12　在分组交换网络中，传输被划分为分组的报文的时间图解

在比较这两个时间图解时，需要注意以下两个事实：

- 在两种网络中，假如传输距离和物理介质是相同的话，信号传播时间t_{prg}具有**相同的值**（same value）。
- 考虑到两种网络的链路带宽是相同的，可能可以得出这样的结论，报文传输时间t_{trns}也会具有**相同的值**（same value）。

但是，将传输的报文切分为分组的过程，以及通过分组交换网络传输这些分组的过程，极大地影响了传送报文到目的节点所需的时间。由于存在额外的延迟，传递时间增加了。

让我们追踪标号为1的单个分组的路由，记录构成传输到目的节点所需总时间的各个部分，然后决定哪一部分是分组交换网络所特有的（图3-13）。

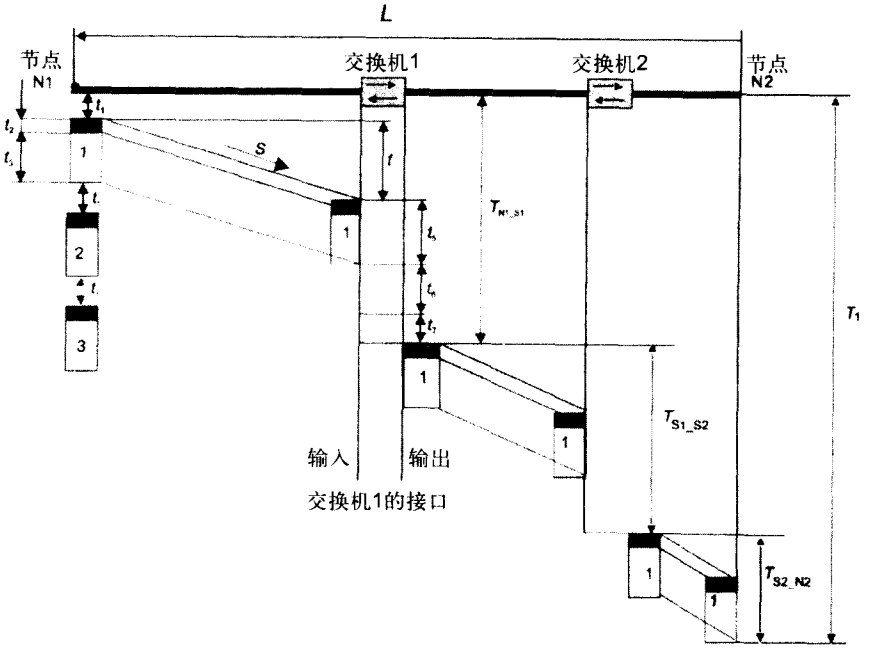


图3-13　在分组交换网络中，传输单个分组的时间图解

从N1节点到交换机1传输单个分组所需要的时间可以由以下各个部分的总和来表示：

- 第一，在源节点，时间延迟由以下部分构成：
  - t_1——生成分组所需要的时间。这一时间也被称为分组生成时间。这一延迟的值取决于发送节点软件和硬件的各种参数，并不取决于网络的参数。
  - t_2——发送分组头部到信道所需要的时间。
  - t_3——发送分组数据域到信道所需要的时间。
- 第二，信号在通信链路上传播需要额外的时间。代表一比特信息的信号从N1节点传播到交换机1所需要的时间表示为t_4。
- 第三，中间交换机也会花费一些额外的时间。这可以用以下部分的总和来表示：
  - t_5——接收分组和它的头部，并放入交换机缓存的时间。正如之前介绍过的那样，这一时间等于（t_2+t_3）——将分组和它的头部从源节点放入链路所需的时间。
  - t_6——分组在队列中花费的时间。这一值可以在很大的范围内波动，并且事先无法得知，因为它取决于当前网络的负载。
  - t_7——将分组交换到输出端口所需的时间。对于特定型号的交换机来说，这一值是固定的，通常相对较小。它可以从几微秒到几毫秒。

将分组从节点N1传输到交换机1的输出接口所需的时间用T_{N1-S1}来表示。这一时间是以下各个部分之和：

$$T_{N1-S1} = t_1 + t_4 + t_5 + t_6 + t_7$$

需要注意的是，t_2和t_3没有出现在这些部分中。从图3-13可以很明显地看到，从发送端到链路的比特传输时间与在通信链路上比特的传输时间是重合的。

在剩下的两个路由部分传输分组所需要的时间分别由T_{S1-S2}和T_{S2-N2}表示。它们和T_{N1-S1}有相同的结构，除了它们不包括t_1这一生成分组的时间。此外，T_{S2-N2}不包括交换的时间，因为在终端节点上中止。因此，在网络上传输单个分组所需要的总时间可以被表示为：$T_1 = T_{N1-S1} + T_{S1-S2} + T_{S2-N2}$。

那么，传输多个分组需要多少时间呢？是传输每个分组所需要的时间之和吗？不是的！让我们回忆一下分组交换网络像流水线一样的工作原理（图3-12）。分组处理发生在多个阶段，所有的网络设备并行完成这些操作。因此，传输这样的一个报文所需的时间会比独立地传输各个分组所需的时间短得多。由于在单个时间点上网络的不确定性，很难精确地计算出这一时间。因此，分组在交换机队列中需要等候的时间也是不确定的。然而，基于分组在队列中等候大致相同的时间间隔这一假设，我们有可能估计出传输具有n个分组的报文所需要的总时间T_{PS}，其值如下：

$$T_{PS} = T_1 + (n-1)(t_1 + t_5)$$

示例　我们使用图3-13所示的例子，对一个分组交换网络中的数据传输延迟的进行了大致估计，并和电路交换网络相比较。假设在两种网络中，需要被传输的文本报文大约有200 000字节。发送端和接收端的距离是5 000公里。通信链路的带宽是2Mb/s.

电路交换网络的数据传输时间由以下部分组成：

- 5 000公里距离的信号传播时间可以大致估计为25微秒。
- 对于给定条件（带宽等于2Mb/s，报文大小等于200 000字节）的报文的传输时间大约为800毫秒。

这意味着传输整个报文所需要的总时间为825毫秒。现在我们来估计用分组交换网络传输同样的报文所需要的额外时间。假设从发送端到接收端的路由包含10台交换机。此外，假设网络并没有处于满负荷状态；因此，在交换机中没有队列。源报文被分为200个分组，每个分组有1 000个字节。

如果我们假设每个分组发送的间隔是1毫秒，那么这些间隔所造成的额外延迟大约是200毫秒。此外，在源节点上会将报文划分为分组，这大约会造成280毫秒的额外延迟。假设分组头部所包含的额外信息大约占报文总大小的10%。因此，传输分组头部所造成的额外延迟大约占了整个报文传输时间的10%（即，80毫秒）。当报文通过每一个交换机时，会产生缓存延迟。对于长度为1 000字节的分组和带宽为2Mb/s的通信链路，对每一个交换机，这一值是4.4毫秒。此外，还存在着交换延迟。在这一例子中，假设交换要花费大约2毫秒。因此，由于缓存和交换，通过10个交换机的分组会带有64毫秒的总延迟。于是，分组交换网络所产生的额外延迟为344毫秒。

对于电路交换网络中825毫秒的数据传输，这一额外的延迟不能算小。虽然这里的计算相当粗略，但是，它帮助解释了为什么对于单个用户，分组交换网络中的数据传输过程通常比电路交换网络中相同的过程慢得多。

基于这样的计算，我们可以得到怎样的结论呢？是不是电路交换网络比分组交换网络更有效率呢？

当考虑网络整体时，使用单个用户流量的传输速度作为效率的标准并不是非常有用。相反地，

使用更加整体性的标准更为合理，例如单位时间内网络传输的数据总量（total amount of data transmitted by the network per time unit）。根据这一标准，分组交换网络的效率被证明比具有相同通信链路带宽的电路交换网络高得多。这一结论在20世纪60年代从实验和分析两个方面都得到了阐述（基于排队理论）。

示例 让我们使用图3-14所示的简单例子，来比较电路交换网络和分组交换网络的性能。两个交换机之间用100Mb/s带宽的链路相连接。用户通过接入链路连接到网络上，接入链路的带宽是10Mb/s。为了简化起见，假设所有的用户以1Mb/s的平均速率生成同样的突发性流量。与此同时，在很短的时间间隔里，这一负载的速率增加到接入链路的最大带宽（即，10Mb/s）。这样的时间间隔持续不超过一秒钟。为了更进一步简化比较，假设所有连接到交换机S1的用户持续性地需要向连接到交换机S2的用户传输信息。

假设图3-14所示的网络是电路交换网络。由于用户流量的峰值达到10Mb/s，因此每一位用户都需要建立一条带宽为10Mb/s的连接。

因此，只能有十位用户同时通过网络传输数据。网络上信息传输的总的平均数据速率为10Mb/s。也就是说，十位用户传输数据，平均每人的速率为1Mb/s。因此，虽然交换机之间通信链路的带宽为100Mb/s，实际上，只有10%的带宽被使用了。

现在，让我们考虑这一网络以分组交换方式进行工作的情况。由于用户流量的平均速率为1Mb/s，所以网络可以同时传输100/1＝100路用户数据信息流，这就充分地利用了连接两个交换机的链路的带宽。但是，这一情况的成立有一个条件——当拥塞发生时，数据流的总速率超过100Mb/s，交换机的缓存容量足够存放分组。让我们试着来估计一下交换机S1所需要的缓存的大小。我们知道每一数据流以最大10Mb/s的速率传输数据（受接入链路带宽限制），并且传输的时间间隔不大于一秒钟，在这一期间，这一流会传输10Mb的用户数据，并且，在网络拥塞最坏的情况下，100路这样的流会到达交换机S1的输入接口，这段时间内，到达交换机S1总数据量为1 000Mb。在同样的时间段内，交换机S1只能够将100Mb的数据传输到输出链路。因此，为了保证在网络拥塞时不会丢失分组，我们需要保证交换机S1最少需要有1 000−100=900Mb的输入缓存。大约换算为100MB。这一存储容量对当代处于工业而言是相当大的。通常，交换机只有较小的缓存容量，在1MB到10MB之间。然而，不要忘记，对于所有的流，峰值负载的时间段同时到达也是小概率事件。因此，即使交换机只有较小的缓存，不足于应付最困难的情况，但在大多数情况下，它能很好地处理网络负载，为每一个流提供更好的服务质量。

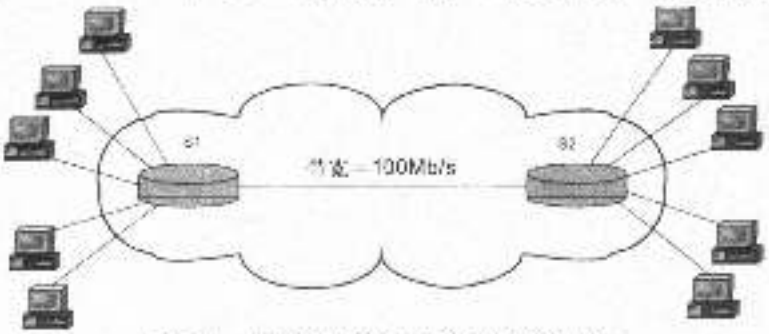


图3-14 分组交换网络和电路交换网络的比较

这里，我们可以用多任务操作系统来进行类比。在这样的系统中，每一个程序或任务都会比在单任务系统中运行更长的时间。在单任务系统中，所有的处理器时间都会分配给这一程序，直

到它结束执行。但是，在多任务系统中，每一时间单位内执行的程序总数要多得多。在单任务操作系统中，或是处理器、或是外部设备总是时不时地空闲，与单任务操作系统类似，电路交换网络在传输突发性流量时，常常不能使用大部分被保留的信道带宽。

分组交换网络不确定的吞吐量是它总体效率提高的代价。当然，单个用户的利益有时候多多少少会有些违背。类似地，在多任务操作系统中，我们不可能去预测应用程序的执行时间，因为这取决于其他应用程序的数量，它和它们必须共享处理器。

作为这部分的总结，让我们来看一下这张表格，它总结了两种网络的性质。基于这些信息，我们可以做出更有根据的决定，什么时候使用电路交换网络更有效率，什么时候使用分组交换网络更好。参见表3-1。

表3-1　电路交换网络和分组交换网络的性质

电路交换	分组交换
在开始传输之前，需要建立连接	建立连接并不是必要的（数据报方法）
只有在建立连接时（连接建立）才使用地址	地址以及其他信息跟随每个分组传递
网路可以拒绝用户建立连接的请求	网络总是可以接收用户数据
保证交互的用户之间的带宽	对单个用户而言，信息速率是不知道的，传输延迟是随机的
实时信息可以无延迟地传输	在传输突发流量时，网络资源可以得到充分利用
传输的高可靠性	由于缓存溢出，可能产生数据丢失
电路带宽的非有效利用减少了网络的整体效率	根据流量的需求，可以在所有用户间自动动态地重新分配物理链路的带宽

3.4　在共享介质网络中的分组交换

在本章的前几节，我们介绍过在几个接口之间共享链路的原理，或者，换句话说，共享介质的原理。现在，让我们来解释一下这些原理是如何在分组交换局域网中工作的。

介质共享曾经一度是最流行的局域网概念：这一原理是一些流行的技术的基础，例如以太网、光纤分布式数据接口以及令牌环。但是，人们也同意基于介质共享的网络已经过了它们最流行的时间。目前，局域网中的主流是交换式以太网。另一个方面，网络世界变换得如此之快，以致于非常明显地，人们重新对共享介质技术出现了兴趣。

共享介质应用的新领域的例子包括：家庭有线网络、个人和本地无线网络。家庭PNA技术出现了，它专门为家庭用户开发。这一技术代表着一个标准以太网技术的改进，家庭电话线和电线被用做共享介质。基于蓝牙技术的个人无线网络用来连接高技术的个人设备（除了台式机之外，还包括PDA、移动电话、高技术电视，甚至是电冰箱），它也使用了共享介质的原理。

此外，无线以太局域网出现了，并且快速地流行起来。这些网络用来在机场、火车站，以及其他移动用户大量聚集的地方，将用户连接到互联网上。但是，没有什么是真正新的，经典的以太网由ALOHA无线网络发展而来，后者由夏威夷大学开发，在这里，共享介质第一次得到了尝试。只是，空气在很长一段时间内并没有作为以太网标准的可用介质，虽然市场上总是有某些公司异乎寻常的产品。随着20世纪90年代晚期无线以太网的来临，历史的合理性才得以恢复。

3.4.1　介质共享的原理

共享介质（shared medium）是用来进行数据传输的物理介质，多个网络的终端节点直接连接到这一物理介质，并且只能轮流使用它。这意味着在任何时候，只有一个终端节点可以访问共享介质，使用它传输分组到另一个连接到这个介质上的节点。

可能的共享介质类型包括同轴电缆、双绞线、光纤和无线电波。

随机访问方法（random access method）是介质共享的一种可能的方法，随机访问方法这一原理也是以太网技术的基础。在这一情况下，访问通信链路的控制是非集中的：所有的网络接口参与这一过程。特别地，在计算机中，**网络适配器**（network adapter）或**网络接口卡**（network interface card）这些特殊控制器提供对共享介质的访问。

以下是随机访问方法的思想：

- 在这样的网络中，计算机只有在介质可获得的情况下，才能通过网络传输数据。介质可获得指，当前计算机之间没有正在进行的数据交换操作，介质上没有电（或者光）信号。
- 在确保介质可获得之后，计算机开始数据传输，这样就**独占**（monopolizing）了介质。提供给单个节点的、用以排他性访问共享介质的时间，由需要传输单个帧所需要的时间来限制。
- 当帧被传输到共享介质上时，所有的网络适配器同时开始接收这一帧。每一个适配器检查被放置在帧的起始域中的目的地址。
- 如果地址与适配器自身的地址一致，那么这一帧被放置到网络适配器的内部缓存中。这样，目的计算机就接收到了要传输给它的数据。

在使用随机访问方法时，可能发生这样的情况：两台或两台以上的计算机同时认定网络空闲，并开始传输信息。这一情况被称做**冲突**（collision），它是网络进行正常数据传输的障碍。这时，多个传输器发出的信号将互相重叠，使最终的信号扭曲。所有基于共享介质的网络技术都提供一个算法，用来检测和正确处理冲突。冲突发生的概率取决于流量的强度。

当检测到一个冲突时，试图传送帧的网络适配器停止传输，暂停一段随机的时间长度，然后再一次试图访问共享介质，重新传输刚才造成冲突的帧。

控制性访问方法（determinate access method）是访问共享介质的另一种方法。这一方法基于使用一种特殊类型的帧，通常被称做**标记**（marker）或**访问令牌**（access token）。计算机只有在持有令牌时才有权力访问共享介质。计算机持有令牌的时间是有限制的，这样，在这一时间段过去后，计算机必须将令牌传递给另一台计算机。

定义令牌传递顺序的规则必须保证，每一台计算机在一个固定的间隔里都可以访问共享介质。

控制性访问方法既可以由集中式方法实现，也可以由非集中式方法实现。对于前一种方法，网络不需要包含任何特殊节点，这些节点定义了访问共享介质的队列；对于后一种情况，存在一个称为访问仲裁器的节点。

3.4.2 LAN结构的理由

第一代局域网由少量的计算机组成（通常是10~30台之间），对所有加入局域网的设备使用单一的共享介质。同时，由于技术的限制，网络有典型的拓扑结构——对于以太网：公共总线（星型）；对于光纤分布式数据接口和令牌环：环型。这些技术具有一致性的特征（即，这类网络中的计算机在物理链路的层次上没有区别）。这种结构的一致性简化了增加计算机数量的过程。它也简化了网络的运行和维护。

但是，在构建大规模网络时，链路的一致性结构成为了一种缺点。在这样的网络中，使用典型的结构成为一种限制，其中最重要的如下：

- 限制了网络节点之间链路的长度
- 限制了网络节点的数量
- 限制了网络节点产生的流量的强度

例如，基于同轴电缆的以太网技术允许使用不长于185米的电缆，不超过30台计算机可以连接到这条电缆上。但是，当计算机开始高强度的信息交换时，它们的数量必须被减少到20，甚至是

10。这保证了每台计算机可以分享总信道带宽的一部分，并且所分享到的带宽是可接受的。

为了消除这些限制，网络基于特殊的通信设备来构造：

- 转发器
- 集线器
- 网桥
- 交换机

3.4.3 LAN的物理构造

我们需要区分物理网络链路的拓扑（**物理网络结构**（physical network structure））和逻辑网络链路的拓扑（**逻辑网络结构**（logical network structure））。

物理链路的配置由计算机之间的电（或者光）连接来定义，可以用图的形式表示，其中，节点是计算机和通信设备，连线代表连接节点对之间的电缆。逻辑链路对应于路由，通过这些路由，信息流在网络上传输。它们创建了通信设备的恰当配置。

在某些情况下，物理网络拓扑和逻辑网络拓扑碰巧相同。例如，图3-15a所示的网络，它具有环型物理拓扑。假设加入这一网络的计算机使用令牌访问方法。令牌总是从计算机到计算机之间顺序传递，并以计算机构成物理环的顺序传递。这意味着，计算机*A*将令牌传送给计算机*B*，计算机*B*将令牌传送给计算机*C*，如此继续。在这一情况下，逻辑网络拓扑是环的拓扑。

图3-15b所示的网络是一个物理网络拓扑和逻辑网络拓扑不一致的例子。在物理上，计算机根据公共总线（星型）拓扑连接起来。但是，访问总线并不遵守以太网技术中的随机访问算法。相反，它通过环型顺序传递令牌来完成：从计算机*A*到计算机*B*，从计算机*B*到计算机*C*，如此继续。这里，传递令牌的顺序并不反映物理链路的顺序。相反，它由网络适配器的驱动程序的逻辑配置来决定。我们也可以用另一种方式配置网络适配器和它们的驱动程序，使得计算机以另一种顺序构成环，例如：*B*，*A*，*C*。但是，不管怎样，物理网络结构没有改变。

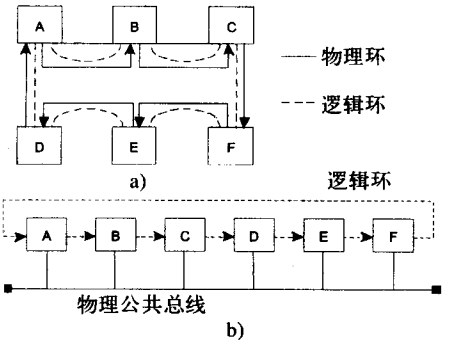


图3-15　网络的逻辑拓扑和物理拓扑

从物理上构造公共共享介质是建立高质量局域网的第一步。物理构造的目的，在于网络可以由几段物理电缆建立，而不必建立在单一的一段电缆上。如果这些可以进行，那么这些物理上不同的段可以像一个公共共享介质那样运作（即逻辑上，它们依然是无法区分的）。

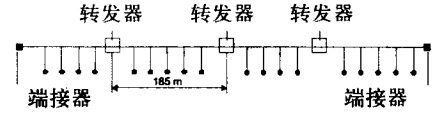


图3-16　转发器使得以太网的长度增加

局域网物理构造的主要方式是转发器和集中器，或称集线器。

转发器（repeater）是一种最简单的通信设备，它用来从物理上连接局域网电缆的不同的段，从而增加网络的总长度。转发器将一个网络段到达的信号重新传输到其他的网络段中（图3-16），同时改善这些信号的物理特性。例如，转发器放大了信号，改善了它的形状和同步性。后者通过修正脉冲之间间隔的不均匀来完成。这样，转发器克服了通信链路长度的限制。由于节点发送到网络的信号流在所有网络段上传播，这样的网络保持了共享介质网络的特性。

具有多个端口，连接多个物理段的转发器常常被称为**集中器**（concentrator）或**集线器**（hub）。从它们的名字反映出，网络段间的所有链路都集中在这一设备上。

要点　*在网络中增加集中器总是会改变网络的物理拓扑，但是它的逻辑拓扑结构不会变化。*

集中器实际上是所有基本局域网技术的必备设备——以太网、Arcnet、令牌环网、光纤分布式数据接口、快速以太网、千兆以太网以及100VG-AnyLAN[3]。

我们需要强调的是，在任何技术中，集中器的运作都有许多相似之处。它们从其中的一个端口转发信号到其他的端口。事实上，需要转发输入信号的端口会有一些不同。因此，以太网的集中器将输入信号转发到除了信号到达端口之外的所有其他端口（图3-17a）。另一方面，令牌环网集中器（图3-17b）从某一端口转发输入信号到另一个端口——即到环中下一台计算机连接的端口。

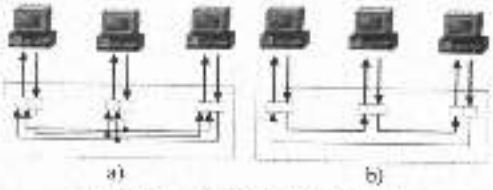


图3-17 不同技术中的集中器

3.4.4 共享介质网络的逻辑构造

网络的物理构造无法克服诸如带宽不足，不能在不同网段使用不同带宽的通信链路等重要问题。这时，网络的逻辑构造可以提供一些帮助。

一个典型的物理网络拓扑（总线型、环型或星型）通过提供给所有网络设备一个单一的、用以进行数据交换的共享介质，限制了所有的网络设备。在大型网络中，这一物理拓扑不适合信息流的结构。例如，在公共总线网络中，任何一对交互的计算机在整个数据交换时间内，独占了公共总线。因此，随着网络中计算机数量的增加，总线成为了瓶颈。

示例 假设一家公司有一个非常简单的以太网，只有一个段（图3-18）。这个企业的所有计算机连接到一条同轴电缆。很明显，用户的数量增加了，网络经常会变得繁忙。因此，用户对于网络应用需要等候更长的时间。除此之外，计算机之间连接线路的长度限制变得愈发明显，因为不可能把所有分配给一个新的工作组的电脑都放置在一起。于是，我们决定使用集中器。图3-18的上部显示了基于物理重构后的网络。现在，我们可以在更大的距离内放置电脑，物理网络结构与公司的管理结构也变得更加一致。但是，与性能相关的问题仍然没有解决。例如，在任何时候，计算机A的用户发送数据到邻近的用户B，整个网络都会被阻塞。这并不令人惊奇，因为，根据集中器的工作逻辑，从计算机A发送到计算机B的帧会被所有网络节点的所有接口转发。这意味着，直到计算机B接收完发送给它的帧，网络上其他的计算机都不能访问共享介质。这一情形之所以会产生，是因为集中器的使用只改变了网络的物理结构，而没有改变它的逻辑结构（图3-18下半部分）。根据这一逻辑结构，信息继续在整个网络上传播，所有的计算机都有相同的权力访问介质，无论它们的位置在哪里。

要解决这一问题，需要放弃对于所有节点使用公共的共享介质。因此，在图3-18考虑的例子中，我们希望保证属于部门的计算机所传送的帧，永远不要离开属于它的网络的部分，除非这些帧是发送到属于另一个部门的机器上。另一方面，只有那些要发送到特定部门节点的帧才需要发送到网络上。这样，在每一个部门的范围内，使用了一个独立的共享介质，这一共享介质由各个部门独占。

要点 在某一个网段的范围内传输仅仅属于特定网段的流量，这被称为流量本地化。确定网络的逻辑构造的过程，便是根据本地流量将网络划分为各个段的过程。

这一网络组织的方法极大地提高了网络的性能，因为当某部门的计算机在交换数据时，其他部门的计算机不再需要等待。除此之外，逻辑构造使得网络不同部分的可获带宽可以有所不同。

[3] 并不是所有列出的技术都保持了它们的重要性。例如，Arcnet和100VG-AnyLAN可以被认为仅仅是早期技术解决方案的例子。

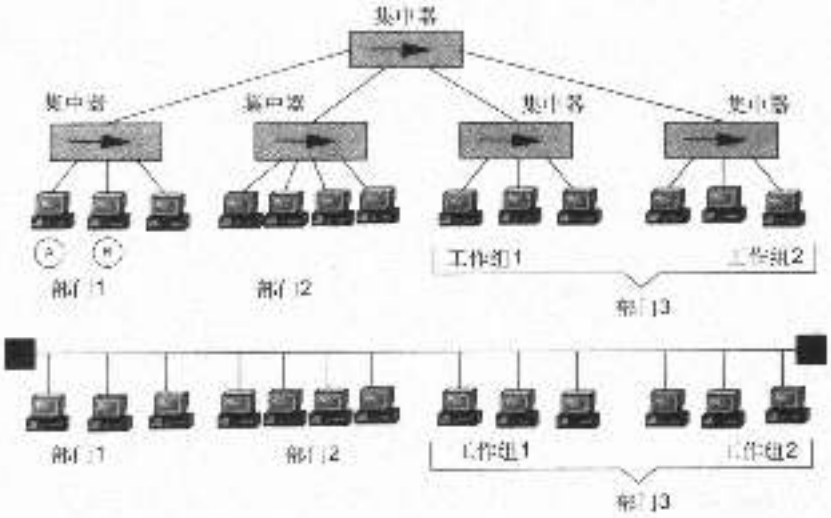


图3-18　网络的物理构造并不提高网络的性能

网络的**逻辑构造**（logical structuring）通过使用网桥、交换机、路由器和网关来实现。

网桥（bridge）将共享介质划分成几个部分（常常称为**逻辑段**（logical segment））。网桥通常只有在必要时，才在段与段之间传输信息来实现这一划分功能，即只有当目的计算机的地址属于另一个段时，才在段与段之间传输信息（图3-19）。通过这一做法，网桥将某个段的流量和另一个段的流量隔离开来，因此，提高了网络的整体性能。流量本地化不但更少地使用带宽，同时还减少了非授权访问数据的可能，因为帧并不离开它们段的范围，入侵者更难截获它们。

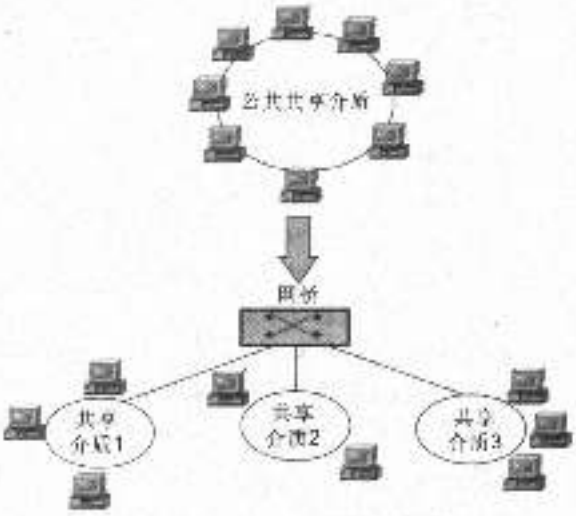


图3-19　网桥划分了共享介质

图3-20中的网络将图3-19中网络的中央集中器替换为网桥。部门1和部门2的网络各含有一个逻辑段，部门3的网络有两个逻辑段。每个逻辑段都基于集中器，并且具有最简单的物理结构——用电缆将所有计算机连接到集中器的各个端口。如果在计算机A上工作的用户发送数据给计算机B的用户，他们在同一个网段中，这一数据只会被限制范围内的网络接口转发。

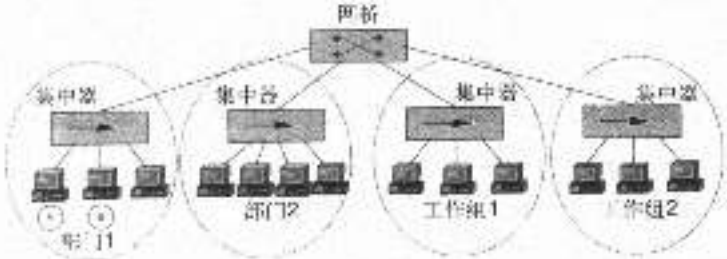


图3-20　网络的逻辑构造：通过使用网桥，公共共享介质被划分为四个独立的共享介质

对于流量本地化，网桥使用了计算机的硬件地址。于是，出现了这样一个问题：网桥怎么知道要将帧转发到哪一个接口？毕竟，硬件地址并不包含特定地址的计算机属于哪个网段的信息。当然，网络管理员可以通过人工设置来告知网桥这些信息。但是，这一方法并不适合大规模的网络。网桥实现了一个简单的学习算法来完成这一任务。

所有到达某一特定接口的帧都是由属于这一网段、连接在这一接口上的计算机产生的。网桥从到达的帧中提取发送端的地址，把它们放在一个特殊的表中，这个表同时包含了特定帧到达的接口。之后，网桥使用这一信息转发帧，如果到目的计算机的路由通过某一接口，网桥便转发帧到这一接口。由于网桥并不知道网络逻辑段间链路的准确拓扑，它只能在段与段之间交互的链路不构成回路的情况下正常工作。

局域网交换机在功能上和网桥非常相似。（在这里，术语“交换机”使用了狭义的语义，专指局域网交换机。）它和网桥的主要区别在于它有更好的性能。交换机的每一个接口都有一个专门的处理器，它使用和网桥中相同的算法来处理帧，并且独立于其他端口的处理器。由于这一特性，交换机的整体性能通常显著优于传统的网桥，后者只有一个的处理单元。可以说，交换机是高级的网桥，以并行的方式处理帧。当在通信设备的每个端口上使用专门的处理器在经济上更有效时，交换机取代了网桥。

3.4.5 作为标准技术例子的以太网

在建造网络的过程中，会遇到许多重要的问题。让我们来看看，解决这些问题的一般方法是如何在基于共享介质的以太网中实现的。基于共享介质的以太网是第一代标准网络技术之一。在本节，我们只考虑一般原理，这些原理构成了共享介质以太网的基础。详细的各种类型的以太网，包括交换以太网，将会在第三部分介绍。

拓扑（*topology*）。以太网标准严格定义了物理链路的拓扑——公共总线（*common bus*）（图3-21）。这张图显示了这种拓扑最简单的实现，所有的计算机连接到公共共享介质，构成单一的段。

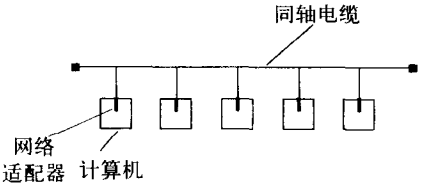


图3-21 以太网

交换方法（*switching method*）。以太网使用数据报分组交换。在以太网中，用于数据交换的数据单元是帧。从功能上说，帧等同于分组。帧有固定的格式，除了数据域外，还包含多种附加信息。

那么，在单一段的以太网中，交换网络在什么地方呢？是不是至少有一个交换机，正如以前提到的那样，是任何分组交换网络的主要元素？或许，以太网代表着一种特殊类型的交换？

事实上，单一段的以太网中确实存在一个交换机，虽然它很难找到，因为它的功能分布在整个网络上。这一以太网的“交换机”由网络适配器和一个共享介质构成。网络适配器是这一虚拟交换机的接口；共享介质担任着交换单元的功能，它在接口之间传输帧。适配器也承担交换单元的功能，因为它们决定着哪一帧是发送到本地计算机的，哪一帧不是。

编址（*addressing*）。每一台计算机——或者，更准确地说，每一个网络适配器——具有唯一的硬件地址。这就是所谓的MAC地址，我们之前曾经介绍过。以太网的地址是扁平的数值地址，因为这里没有使用层次结构。它支持以下类型的地址：单播、广播和多播。

介质共享（*medium sharing*）和多路复用（*multiplexing*）。终端节点使用单一的共享介质进行数据交换，使用随机访问方法。

从以太网终端节点来的信息流以基于分时的方式多路复用到单一的传输链路。这意味着，属于不同流的帧轮流访问链路。为了强调多路复用和介质共享这两个概念的不同，让我们考虑这样的情形，连接到以太网的计算机中，只有一台需要传输数据，这些数据通常由多个应用程序产生。

在这一情况下，在网络接口间共享介质这一问题并不会出现；通过一个公共链路传输多路信息流的问题（即，多路复用问题）仍然存在。

编码（*encoding*）。以太网适配器以20MHz的时钟频率工作，往介质传输矩形脉冲，代表了二进制的1和0。当帧传输开始时，所有的比特都以10Mb/s的恒定的速率传输到网络上。每一比特的传输占据两个时钟周期。这一速率是以太网的链路带宽。

可靠性（*reliability*）。为了提高数据传输的可靠性，以太网使用标准技术。这些技术包括计算校验和，并在帧尾进行传输。如果接收的适配器通过重新计算校验和发现帧数据中存在错误，那么，这一帧就会被丢弃。以太网协议并不重新传输这一帧，这一任务由其他的技术完成（例如，TCP/IP网络中的TCP协议）。

半双工传输方法（*half-duplex transmission method*）。以太网共享介质是半双工通信信道。事实上，网络适配器不能同时使用这一信道进行发送和接收数据。这些任务必须轮流完成。

队列（*queue*）。在基于共享介质的以太网中，可能刚开始看上去，没有分组交换网络中队列的特性。但是，缺少带有缓存的交换机并不意味着没有队列。在这一情况下，队列被搬到了网络适配器的缓存中。当介质忙于传输其他网络适配器的帧时，数据（负载）持续地到达网络适配器。因为在那个时候，数据不能被传送到网络上，它们开始在以太网适配器的内部缓存中积聚，因此形成了队列。正因为这样，和其他所有分组交换网络相似，以太网也存在着帧递送的不同的延迟。

但是，以太网也有它独特的特性。共享介质是一种帧传输速率的调节器，当它繁忙时，它不再接收其他的帧。因此，当网络负载过重时，网络实现了背压，因此强迫终端节点降低传送数据到网络的速率。

小结

- 在电路交换网络中，根据用户的要求建立一些持续性的信息信道，它们被称为电路。在数据传输时，保留了一系列连接用户的通信链路形成了电路。在电路的整个长度上，电路以固定的速率传递数据。这意味着，电路交换网络可以保证延迟敏感数据（语音和视频）的高质量传输，这些延迟敏感数据也被称为实时流量。但是，无法动态重新分配物理链路的带宽是电路交换网络的一大缺陷。这一缺陷使得在传输计算机网络中典型的突发性流量时，这类网络的效率不高。
- 在使用分组交换时，源节点将需要传送的数据划分成小的部分，称为分组。分组带有头部，表明了目的地址。因此，它可以独立于其他数据由交换机处理。在传送突发性流量时，分组交换的方法提高了网络的性能，因为在处理大量独立的流时，这些流的活动周期并不总是相同的。分组被传送到网络上时，并没有根据源生成数据的速率，事先保留资源。但是，这一交换方法也有它的不足：在本质上，传输延迟是随机的，因此，在传输实时流量的过程中，可能会出现问题。
- 分组交换网络可以使用三种转发算法之一：未建立连接的（无连接），也被称做数据报传输；面向连接的以及虚电路。
- 共享介质是数据传输的物理介质（同轴电缆、双绞线、光纤或无线电波），一定数量的网络终端节点直接连接到其上，并且只能以轮流的方式使用共享介质。共享介质原理构成了一些著名技术的基础,例如以太网、光纤分布式数据接口、令牌环。虽然基于共享介质的网络可能已经过了它们最流行的时候，但是，依然有明显的迹象表明人们对这一技术重新燃起了兴趣。例如，诸如家庭有线网络和个人本地无线网络，这些很新的技术使用了共享介质原理。

复习题

1. 电话网络中使用了什么类型的多路复用和交换机制?
2. 电路交换网络的什么性质可以被认为是它的缺点?
3. 分组交换网络的什么性质对传输多介质信息产生了负面影响?
4. 电路交换网络使用缓存吗?
5. 电路交换网络的什么元素可以拒绝要求建立电路的节点?
   A. 没有，网络总是可以从用户处接收数据
   B. 任何中间节点
   C. 目的节点
6. 哪些概念是以太网技术的特点?
7. 数据报网络是否考虑那些已有的流?
8. 请给出逻辑连接的定义。
9. 是否可能在两个终端节点间不建立逻辑连接的情况下，提供可靠的数据传输?
10. 哪些逻辑连接可以被命名为虚电路?
11. 哪些网络使用虚电路技术?
12. 请指出以下列出的设备中，哪些在功能上非常相似:
    A. 集线器
    B. 交换机
    C. 集中器
    D. 转发器
    E. 路由器
    F. 网桥
13. 请列出网桥和交换机的区别。
14. 以下的陈述是否正确? 在中央有集线器的显型拓扑的以太网，要比建立在同轴电缆上、公共总线拓扑的相同的网络具有更高的可靠性。
15. 你如何在基于集线器建立的网络上，增加每台终端用户计算机的可用带宽?

练习题

1. 根据以下给出的数据，和电路交换网络相比，分组交换网络中数据传输时间会增加多少:
   - 总的传输数据的大小——200KB
   - 总的连接链路的长度——5 000公里
   - 假设信号的传播速度——0.66倍光速
   - 链路带宽——2Mb/s
   - 分组大小（不包含头部）——4KB
   - 头部大小——40字节
   - 分组之间的间隔——1msec
   - 中间交换机的数量——10
   - 每个交换机的交换时间——2msec

   假设网络处在低负载模式中，因此，交换机中没有队列。
2. 如果在图3-22中所示的网络段中，所有的通信设备都是集中器，那么，从计算机*A*发送到计算机*B*的帧会出现在哪些端口上?

A. 5和6
B. 4，5和6
C. 4，5，6和7
D. 4，5，6，7和12
E. 所有的端口

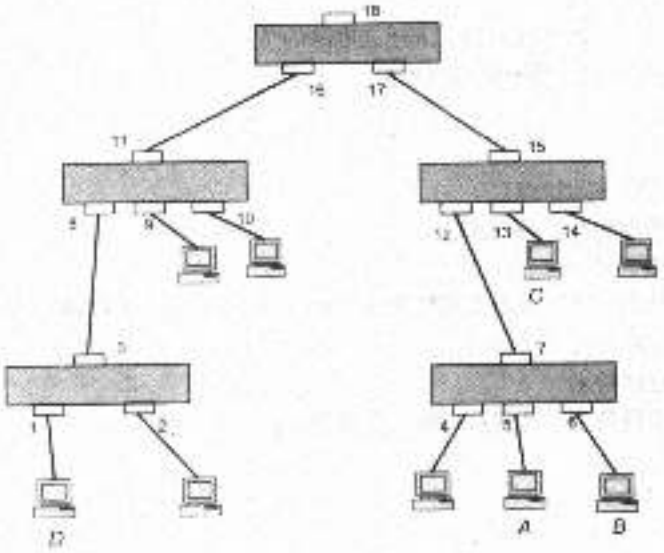


图3-22　网络的段

3. 假如在图3-22所示的网络段中，所有的通信设备都是交换机，那么，从计算机*A*发送到计算机*B*的帧会出现在哪些端口上？
4. 假如在图3-22所示的网络段中，除了将计算机*A*和计算机*B*连接起来的设备是集中器之外，其他所有的设备都是交换机，那么，从计算机*A*发送到计算机*D*的帧会出现在哪些端口上？
5. 在数据报网络中，在节点*A*和节点*B*之间有3条流和3条可用的路由，是否可能通过每一条不同的路由发送每一条流？
6. 在虚电路网络中，在节点*A*和节点*B*之间有3条流和3条可用的路由，是否可能通过每一条不同的路由发送每一条流？
7. 如果一个基于共享介质的网络有10Mb/s的带宽，由100个节点组成，在这样的网络中，两台计算机之间最大的数据交换速率是多少？
8. 一个网络可以用两种模式传送数据：数据报和虚电路。当你选择一种特定的模式传送你的数据时，如果主要的考量是速度和数据传输的可靠性，那么你会考虑选择哪一种方式？
9. 你是否认为电路交换网络很快会被分组交换网络所替代？或者相反，分组交换网络将被电路交换网络所替换？或者是，两种技术将共存？请提供一些理由来说明你的观点。请考虑这些技术不同的应用范围。

第4章 网络体系结构与标准

4.1 引言

网络体系结构是由各种元素构成的网络系统的表示，其中每一个元素完成一项特定的功能。所有的网络元素共同协调它们之间的运作，以便在多台计算机间解决一个共同的交互任务。换句话说，网络体系结构将一个公共的问题分解成一系列子问题，每一个子问题由一个单独的网络部件来解决。网络体系结构中最重要的元素之一是**通信协议**（communications protocol）。这可以被定义为网络节点间交互的一套形式化的规则。

开放系统互连（Open System Interconnection，OSI）的开发是计算机网络体系结构标准化过程中的一个突破。这个在20世纪80年代早期开发的模型，总结了那个时候积累下来的所有经验。OSI模型是一个国际化的标准，定义了一种纵向分解计算机交互问题的方法，这一方法把这个任务交给通信协议来完成，通信协议又被分成七层。通信协议的层次构成一种层级结构，也被称为**协议栈**（protocol stack），其中，每一层使用下面一层，来便捷地解决任务。

目前使用的协议栈（或者到现在还很流行的那些协议栈）普遍反映了OSI模型的体系结构。但是，和OSI体系结构相比，每一个协议栈都有它特定的特点和不同之处。因此，最流行的TCP/IP栈由四层组成，而不是七层。

计算机网络的标准体系结构还决定了网络元素间协议的分布，例如终端节点（计算机）和中间节点（交换机和路由器）。中间节点只支持有限的协议栈功能中的一个子集；它们通过在终端节点之间传送网络流量，来实现传输功能。终端节点支持整个协议栈，因为它们必须提供诸如Web服务的信息服务。这种功能的分布把智能网络功能移到了网络的边缘。

4.2 网络节点互动的分解

在网络设备间组织交互是一件复杂的工作。最常见、著名、通用的解决这一问题的方法是**分解**（decomposition）（即，将一个复杂的问题划分成几个更简单的任务或模块）。分解明确了每个模块严格的功能定义，以及它们之间交互的方式。这被称为模块间的接口。在使用这一方法时，通过将每个模块从它们的内部机制中抽象出来，只注意它们互相交互的方式，每个模块都可以被认为是一个**黑盒**（black box）。通过在逻辑上简化这一问题，我们可以独立地开发、修改和测试每一个模块。因此，图4-1所示的每一个模块都可以在无需修改其他的模块的条件下被重写。让我们考虑模块*A*。假设开发人员保持模块间的接口不变（在这里，这些是*A-B*，*A-C*和*A-F*接口[⊖]），其他模块不需要进行任何改动。

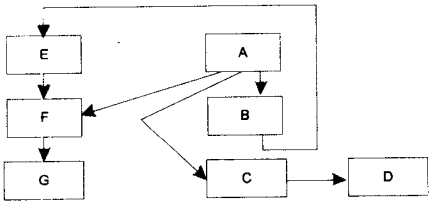


图4-1 分解问题的例子

4.2.1 多层的方法

多层的方法（multilayer approach）是一个更有效的概念。在把原始的任务表示为一系列模块后，这些模块根据层而组织并排序，构成了层级结构。对每个中间层使用层级原理，我们可以

⊖ A-F接口原书中遗漏，由译者加上。——译者注

直接指出在它之上和之下的相邻的层（图4-2）。

构成每个层的模块彼此执行它们的任务时，必须仅仅从其下面层的模块中请求服务。它们只能将它们执行的结果传递到它们之上的那一层的模块。这样的层级分解不但对特定的模块，而且对每一个层都给出了清晰的功能和接口定义。

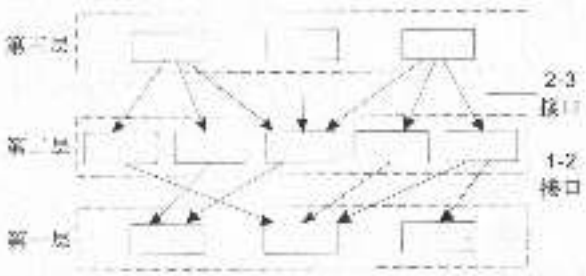


图4-2　多层的方法——创建层级化任务

层间接口（interlayer interface）也被称为**服务接口**（service interface），定义了下面一层对相邻上一层提供的功能集。（参见图4-3）

这一方法允许独立于其他层而开发、测试和修改特定的层。通过从低层次到高层次，层级分解可以创建更抽象的，因此也是更简单的，对系统问题的表示。

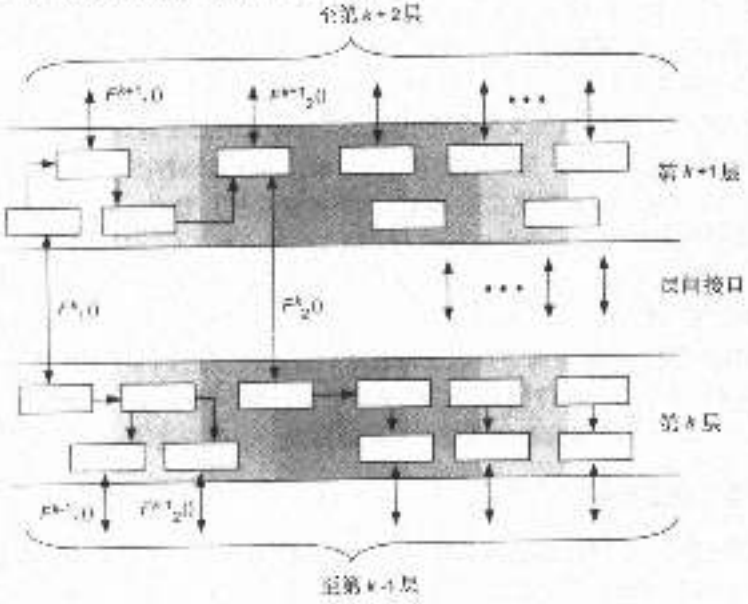


图4-3　多层间交互的概念

示例　让我们考虑一个简化的任务描述——从存储在磁盘上的文件中读取一条逻辑记录。这一任务可以表示为以下特定任务的层级：

1. 使用字符文件名，寻找需要访问数据的文件的特征。这些可以是文件的物理位置，它的大小等等。由于这一层的功能仅仅和目录查找相关，在这一层中，文件系统的表示相当抽象。文件系统表示为一棵树，它的节点是目录，叶子代表文件。这一层并不关心任何其他的硬盘物理和逻辑数据组织的细节。

2. 接下来，我们需要确定文件将要被读取的特定部分。为了完成这一任务，我们需要到文件系统更低的抽象层次。这一层的功能将把文件解释为一系列物理磁盘块，互相以一种特定的方式联系起来。

3. 最后，从磁盘读取需要的数据。在决定物理块的数量后，文件系统要求输入/输出系统完成读操作。在这一层次上，已经需要处理文件系统的细节，例如簇、磁道和扇区的数量。

例如，在请求最高层的文件系统调用应用程序的功能中，可能会类似这样操作：从名字为DIR1/MY/FILE.TXT的文件中读取第22条逻辑记录。

高层次的抽象无法仅仅靠它自己完成这一请求。在通过字符名称（DIR1/MY/FILE.TXT）定义完物理地址后，它发送以下请求到下一层：读取位于以下物理地址的文件的第22条逻辑记录：174，大小为235。

作为对这一请求的回应，第二层确定地址为174的文件在磁盘上有五个非邻接的区域，所要求的记录位于文件的第四个部分，位于物理块345。然后，它向磁盘驱动发送请求，读取所需的逻辑记录。

根据我们的简化方法，文件系统层级间的交互是单向的，从顶到下。但是，真实的情况要复杂得多。为了确定文件的特性，最高层必须对文件的字符名称进行解码（即，在全部文件名中，顺序读取整个的目录结构）。这意味着，最高层需要向下层不止一次地发送请求。下层必须数次请求磁盘驱动读取物理磁盘的目录结构数据。每一次所完成操作的结果必须从底部传回到顶部。

通过使用网络在计算机之间组织交互的问题，也可以被表示为一系列按层级组织的模块。例如，在相邻节点间保证可靠数据传输的任务可以被分配给更低的层次。更高层次的模块可以负责在整个网络中的消息传输。显然，后一个任务——在任何两个网络节点间组织交互，而并不仅仅是相邻的节点——更为普遍。因此，这一问题可以通过使用对下一层的多次请求来完成。因此，节点*A*和节点*B*之间交互的组织（图4-4）可以被简化为：顺序连接中间节点对。

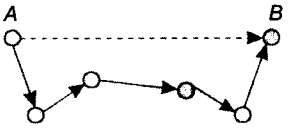


图4-4　连接任何的网络节点对

4.2.2　协议和协议栈

由于在报文交换的过程中，至少需要涉及**两方**或两方以上，因此，网络的多层表示工具有其特定的特点。这意味着，在这个特定的情况下，需要组织两个网络工具层次进行协调良好的运作，而这两个工具分别运行在两台计算机上。网络交换参与的双方都必须接受一些协定。例如，它们必须就电信号的水平和形式、就报文大小的方法、以及差错检测的方法达成一致。换句话说，在所有的层上都必须达成协议，从最底层——比特传输层——到为网络用户实现服务的最高层。

图4-5表明了两个节点间交互的模型。每一方的互连工具都有四层。每一层支持两类接口。首先，对网络工具本地层级的上层和下层有服务接口。其次，还必须有对另一方交互工具的接口，它们在同样的层上。这种类型的接口称为**协议**（protocol）。因此，协议总是**对等接口**（peer-to-peer interface）。

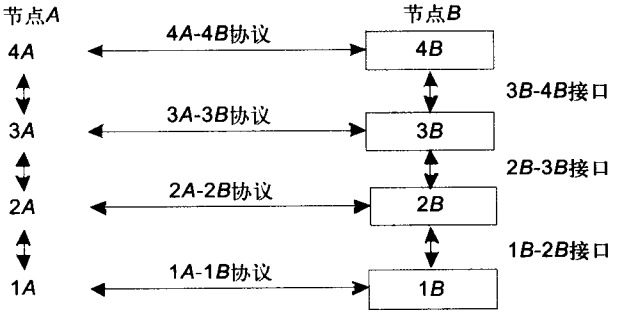


图4-5　两个节点间的交互

要点　大体上，术语协议和接口表示同一样东西，也就是，两个物体间交互过程的标准

化描述。但是，就传统意义而言，在网络中它们有不同的应用领域：协议定义了不同节点上运行的同一层模块间的交互规则，而接口定义了在同一个节点上相邻层模块间的交互规则。

一个能够实现网络节点交互的、层级化组织的协议集被称为**协议栈**（protocol stack）（或者是**协议集**（protocol set）或**协议组**（protocol suite））。

低层次的协议通常由硬件和软件共同实现；更高层次的协议只由软件实现。

实现一个特定协议的软件模块被称为协议实体，或简单地称为协议。对同样的协议，实现的效率会有差别。这就是为什么在比较不同的协议时，需要考虑协议的运作逻辑和协议实现的质量。此外，网络设备间交互的效率取决于构成协议栈的整套协议的质量。特别是，我们需要衡量当功能被分布到不同层的协议中时，效率是否高，协议层之间的接口定义得是否清晰。

交互双方的同一层协议体根据协议规则交换报文。通常，报文包含头部和数据域（有时候，它也可以被留成空白）。报文交换是一种语言，双发需要互相解释给对方，在交互的每个阶段，哪些工作需要完成。每一个协议模块的运作包括解释到达报文的头部，并完成相应的操作。不同协议的报文头部有不同的结构，对应于它们在功能上的差别。报文头部的结构越复杂，分派给相应协议的功能就越复杂。

4.3 OSI模型

协议是交互的网络节点双方接受的一个约定，但是，这并不能证明这一协议是一个标准。实际上，网络架构师在实现网络时，总是不遗余力地使用标准协议。这些协议可能是私有的、全国性的，或者是国际的标准。

在20世纪80年代早期，几个专门进行国际标准化的组织，包括**国际标准化组织**（International Organization for Standardization，ISO）和**国际电信联盟电信部**（ITU Telecommunication Standardization Sector，ITU-T），开发了**开放系统互连**（Open Systems Interconnection，OSI）。这一模型在后来计算机网络的发展中发挥了重要的作用。

4.3.1 OSI模型的一般特性

到20世纪70年代晚期，那时有大量专用的通信协议栈，其中的例子包括DECnet和系统网络架构（SNA）。这种互连工具的多样性使得使用不同协议的设备很难互相兼容。那个时候，克服这一问题的可能的方法之一是转化到使用统一的、公共的协议栈，这些协议栈是为了克服现有协议栈的缺点，而被创建出来的。这一学术化的开发新协议栈的方法来源于OSI模型的开发。OSI模型不包含任何特定协议栈的描述，因为它的目的不同——为了描述一个通用的网络互连工具。OSI模型作为一种网络专家的通用语言被开发出来。正因为这样，它常常被称作**参考模型**（reference model）。OSI模型的开发花费了七年时间（从1977年到1984年）。

OSI模型定义了以下内容：

- 分组交换网络中系统各个层之间的相互通信
- 这些层的标准名称
- 每一层必须完成的功能

在OSI模型中（图4-6），相互通信的工具被划分为七个层次：应用层、表示层、会话层、运输层、网络层、数据链路层和物理层。每一层处理严格定义了网络互连的一个方面。

要点　OSI模型只描述了由操作系统实现的系统工具、一些系统的设施，以及系统的硬件。这一模型没有包含用户应用程序之间交互的工具。区分应用程序的交互和OSI模型中的应用层之间的区别非常重要。

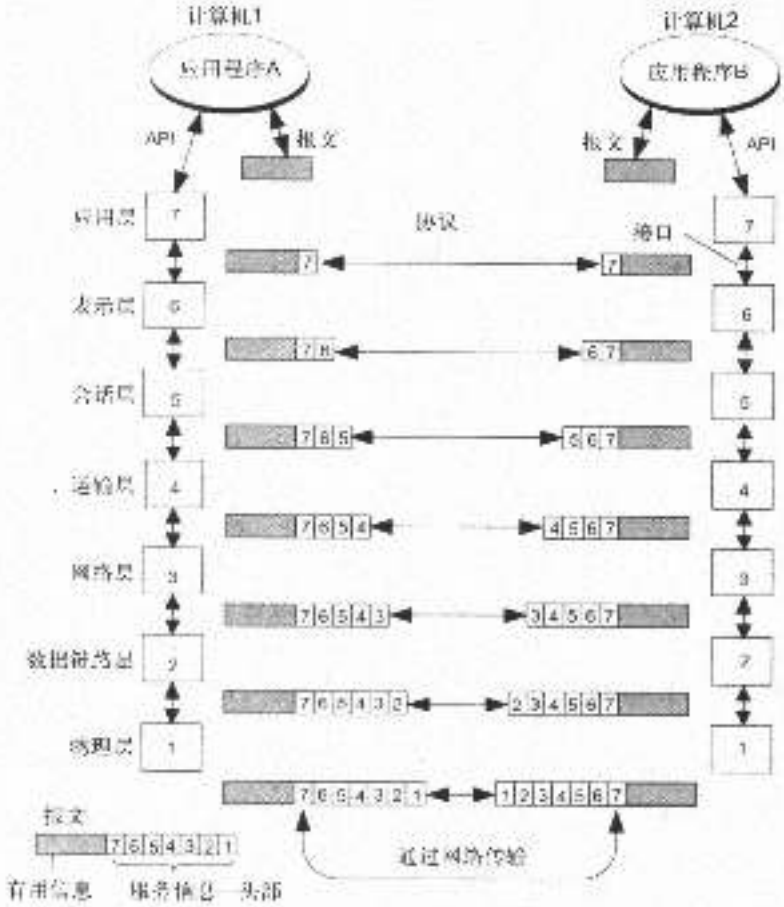


图1-6 ISO/OSI模型

应用程序可以使用这套七层的系统工具，实现它们自己的交互协议。为了这一目的，一套特殊的**应用程序编程接口**（application programming interface，API）被提供给了程序员。根据权威的OSI模型的设计，应用程序可以将它的请求发送给最高的层——应用层。但是，事实上，大多数的通信协议栈提供给程序员直接调用低层服务。

例如，有些数据库管理系统有内置的远程文件访问的工具。这时，在访问远程资源时，应用程序并不使用系统的文件服务，相反，它绕过OSI模型的上面那些层，直接请求系统工具进行报文传输，这些工具处在OSI模型的较低的那些层。

假如运行在计算机1上的应用程序A需要和运行在计算机2上的应用程序B通信。为了实现这一目的，应用程序A请求一个应用层服务，例如，文件服务。基于这一请求，应用层软件以标准的格式形成了一个报文。但是，为了将这一信息递送到目的地，还需要完成几项其他的任务，它们被指派给了下面各层。

在形成了报文之后，应用层将它向下传递给表示层。基于从应用层报文头部（header）接收到的信息，表示层协议完成了需要的动作，将它自己的信息——表示层头部——加到报文中。表示层头部包含了针对目标机器表示层协议的指令。产生的报文再被传送到下面的会话层，它也加上了自己的头部，如此继续。有些协议的实现不但将它们自己的信息放在报文的开始，也就是头部，而且还放在报文的末端，也就是所谓的**尾部**（trailer）。最后，报文到达最低的物理层，物理层负

正通过连接的链路将报文传送到目标机器。这个时候，这一报文带有所有层的头部（图4-7）。



图4-7　不同层报文的嵌套

物理层将报文放到计算机1的输出接口上，从那里，报文开始在网络上传输。请注意在这之前，报文只是在计算机1的内部一层到一层之间传输。

当报文被送送到计算机2时，它由计算机2的物理层接收，然后顺序地一层一层地向上传递，每一个层检查它自己那一层的头部，完成所要求的功能，然后删除头部，将报文交给上一层。

很明显，同一层的协议体从不直接通信。这一交互总是通过下面层协议的工具来协调。只有不同节点的物理层会进行直接交互。

除了术语**报文**（message）之外，网络专业人士还使用其他的术语描述被数据交换的单元。在ISO标准中，不同层协议的数据交换单元有一个通用的名称——**协议数据单元**（Protocol Data Unit，PDU）。专门的术语用来描述不同层的数据交换单元：**帧**（frame）、**分组**（packet）、**数据报**（datagram）以及**段**（segment）。

4.3.2　物理层

物理层（physical layer）通过使用诸如同轴电缆、双绞线、光纤，或者长距离数字电路这样的物理链路，处理字节流的传输。我们已经在第2章中讨论了这一层的基本性质（参见2.3节）。

物理层的功能在所有连接到网络的设备中实现了。就计算机而言，物理层的功能或是通过网络适配器完成，或是通过串口完成。

以太网技术的10Base-T标准是物理层协议的一个例子。这一标准定义了电缆为带有100欧姆阻抗的非屏蔽三类双绞线、RJ-45接口、最大段长度100m，线上的数据表示为曼彻斯特码，以及其他一些传输媒体和电信号的特点。

物理层并不关心它传输的信息的意义。从它的角度看，这些信息是需要根据指定的时钟频率（预定义的相邻比特之间的间隔），无损传输到目的地的比特流。

4.3.3　数据链路层

数据链路层（data link layer）是工作在分组交换模式下的第一层（从下往上数）。在这一层中，PDU通常被称为**帧**（frame）。

对于局域网和广域网，定义了不同的数据链路层的功能。当OSI模型还在构造中时，局域网和广域网的技术相差甚远，以致于无法将它们的运作无条件地概括出来。因此，数据链路层的工具必须提供如下的功能：

- **在局域网中**　保证任何（*any*）两个网络节点间帧的递送。我们假设网络有一个典型的拓扑，例如公共总线、环型、星型，或者树型（层级星型）。使用被限制在标准拓扑的网络的例子包括：以太网、光纤分布式数据接口、令牌环网。

- **在广域网中**——在两个通过单一通信链路连接的相邻节点中，保证帧的递送。**点对点协议**（point-to-point protocol）（正如这些协议被称做的那样）的例子包括广为人知的PPP协议和HDLC协议。基于点对点的链路，可以建立任何拓扑结构的网络。

对于互连局域网，或者对于保证广域网中任何两个节点间帧的递送，就需要使用更高层的网络工具。

数据链路层实现的功能之一是提供接口给下面的物理层和更高的一层（网络层）。网络层将需要网络传输的分组发送给数据链路层，并且接收从网络到达的分组。数据链路层使用物理层，或是从网络接收比特序列，或是将比特序列传送到网络上。

让我们考虑数据链路层的操作，从发送端的网络层与数据链路层通信，并将带有目的节点地址的分组发送给数据链路层开始。为了完成这一任务，数据链路层创建一个帧，这一帧包含数据域和头部。数据链路层将分组封装进帧的数据域，并且在帧的头部添上了适当的服务信息。要被网络交换机转发的分组所使用的目的地址是帧的头部中最重要的信息。

差错检测和纠正是数据链路层的另一项任务。为了实现这一目标，数据链路层通过在帧的头部和尾部放置一系列特殊的比特，修改了帧的边界。之后，数据链路层在帧上增加了一个特殊的校验和，也称为**帧校验序列**（Frame Check Sequence，FCS）。这一校验和根据一个特定的算法，由所有组成帧的字节计算出来。使用这一FCS值，目的节点可以确定帧中数据在网络上传输的过程中是否产生了差错。

但是，在通过网络将帧传送到物理层以供传输之前，数据链路层可能需要解决另一个重要的问题。如果网络使用共享介质，那么，在物理层开始数据传输之前，数据链路层必须检查介质的可用性（没有使用共享介质时，这样的检查可以被忽略）。实现共享介质可用性检查的功能有时候被分为一个独立的子层——**介质访问控制**（Medium Access Control，MAC）。

如果共享介质已经被释放，物理层将帧传送到网络上，帧通过通信链路传送，然后以比特序列的形式到达目的节点的物理层。这一层将接收到的比特向上传送到目的节点的数据链路层。数据链路层将比特组合成帧，重新计算接收到的数据的校验和，并且将结果和帧的校验和相比较。如果校验和的值相符，那个这个帧就被认为是正确的。如果校验和的值不相符，就报告一个错误。通过重新传输出错的帧，数据链路层的功能既包括差错检测，也包括差错纠正。然而，这一功能并不是必备的。有些数据链路层的实现，例如以太网、令牌环网、光纤分布式数据接口、帧中继，没有这一特性。

数据链路层的协议由计算机、网桥、交换机和路由器实现。在计算机中，数据链路层的功能由网络适配器和它们的驱动程序共同协调实现。

数据链路层协议通常在一个大型网络的一部分中工作，由网络层的协议将它们连接起来。数据链路层的地址仅仅在一个网络内部用以传输帧；更高层（网络层）的地址用来在网络之间转发分组。

在局域网中，对于网络节点间报文的转发，数据链路层的功能相当强大，并且提供了完整的一套功能。在某些情况下，局域网的数据链路层协议是自给自足的运输工具，可能会允许应用层协议或者应用程序直接在它们之上进行操作。这时，就没有必要使用网络层或运输层的工具了。然而，为了保证对于任何拓扑，网络中报文的高质量传输，数据链路层的功能并不够。对于广域网，这尤其明显，在广域网中，数据链路层协议实现了最邻近节点间的简单数据传输功能。在OSI模型中，报文高质量传输的任务被指派给了相邻的上两层——网络层和运输层。

4.3.4 网络层

网络层（the network layer）将多个网络连接起来，创建了统一的运输系统，也被称为**互连**

网络（internetwork），或者简称为互连网（internet）。请不要将术语*internet*和*Internet*混淆起来。后者是建立在TCP/IP技术基础上、最著名的互连网络的实现，这一网络覆盖整个世界。

在多个网络间连接多网络连接起来的技术称为网络互连（internetworking）。这一技术通常建立在许多技术之上。

图4-8中有多个网络，每一个都使用了一种特定的数据链路技术：以太网、光纤分布式数据接口、令牌环、ATM、帧中继。基于这些技术，每一个网络都可以在本地网络中将两个用户连接起来，但是一个网络不能传输数据到另一个网络。对此，原因非常明显：网络技术的显著差别。即使对于最相似的局域网技术——以太网、光纤分布式数据接口、令牌环网——实现相同的编址系统（MAC子层地址，也称为MAC层址）也存在帧格式的差别、协议运行逻辑的差别。局域网和广域网的技术有更多的差别。大多数广域网技术使用之前建立的虚电路，它们的标识符被用作地址。所有的技术使用特定的帧格式，ATM甚至有专门的术语来指代它们——信元（cell）。当然，它们也都有它们自己的协议栈。

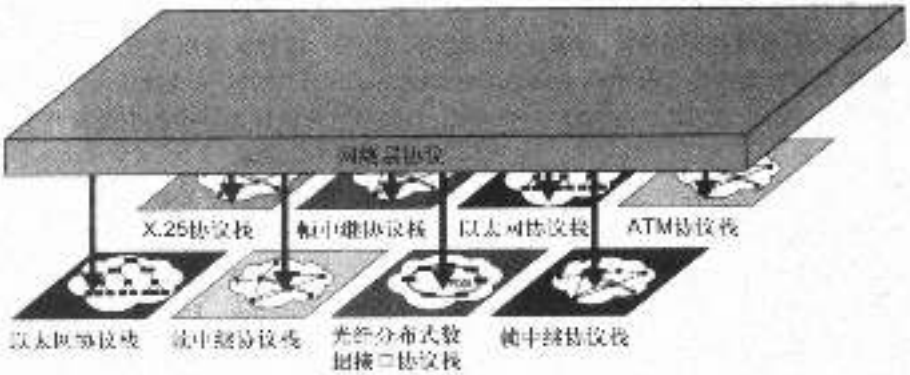


图4-8　网络层的必要性

基于如此相差甚远的技术连接网络，我们需要额外的工具。OSI模型中的网络层提供了这种工具。

网络层的功能由以下实现：

- 一组协议
- 称为路由器的特殊设备

保证不同网络的物理连接是路由器（router）的功能之一。路由器有多个类似于计算机接口的网络接口，其中每一个都连接到网络上。因此，所有的路由器接口都可以被认为是不同网络的接口。路由器可以基于通用计算机，由软件模块来实现。例如，Unix和Windows的典型配置中包含了一个软件路由器。然而，通常路由器由专门的硬件平台实现。路由器软件包含网络层的协议栈。

因此，为了互连图4-8中的网络，需要使用路由器连接所有的这些网络，并且在所有需要使用互连网络通信的终端计算机上安装网络层协议栈（图4-9）。

需要通过互连网络传输的数据从更高的运输层到达网络层，然后它们被加上了网络层的头部。数据和头部构成了分组——指代网络层PDU的常用术语。网络层分组的头部有统一的格式。这一格式并不取决于数据链路层帧的格式，这些帧的格式可能针对互联网络中特定的网络。除其他信息之外，这一头部还包含了分组的目的地址。

为了保证网络层协议能将分组传递给互联网络的任何节点，我们需要保证每一个节点都有一个在整个互联网络范围中唯一的地址。这样的地址称为网络地址（network address）或全局地址

(global address)。需要和其他节点交换数据的互连网络的每一个节点都有一个网络地址。同时还有数据链路层技术分配给它的地址。例如，在图4-9中，互联网络中以太网内的计算机有数据链路层地址*MAC1*，以及网络层地址*NET-A1*。同样，在*ATM*网络中，由虚电路*ID1*、*ID2*标识的节点有网络地址*NET-A2*。网络层的分组必须标识网络层的地址作为目的地址。分组的路由将根据这一地址而决定。

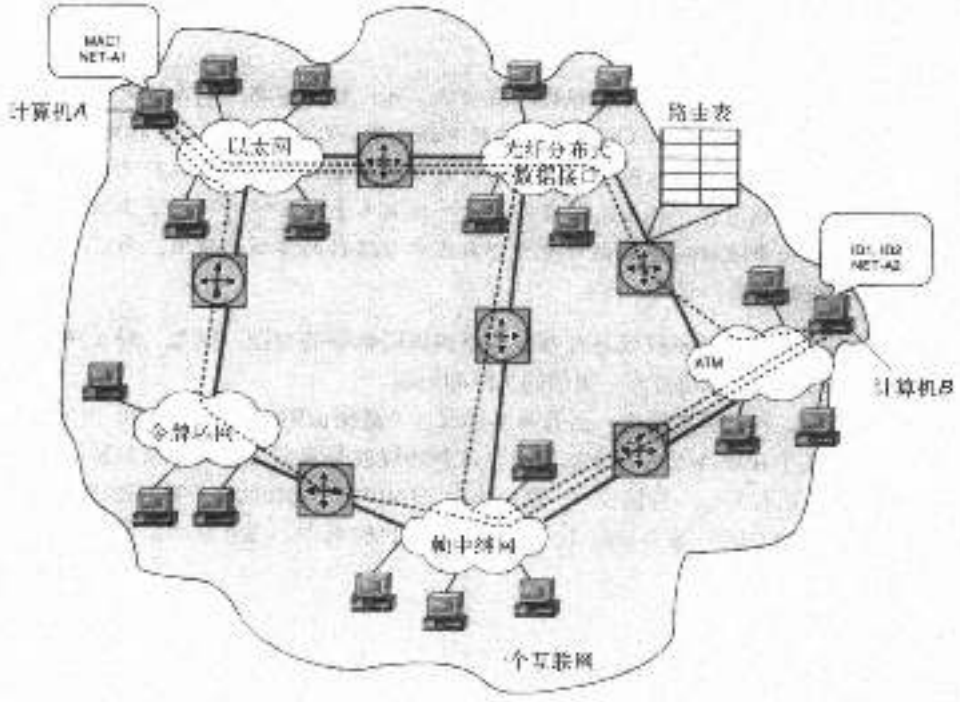


图4-9 互联网络的例子

路由（routing）是网络层的重要任务。路由由分组通送到目的节点的过程中，通过的一些网络（或路由器）来描述。图4-9指出了数据从计算机*A*传输到计算机*B*的三条路由。路由器收集互联网络链路的拓扑信息，根据这一拓扑创建交换表，请注意在这个时候，这些交换表有特殊的名称——**路由表**（routing table）。第2章简单介绍了选择路由的方法（参见2.5节）。

根据多层的方法，网络层寻求下面的数据链路层来实现这一任务。通过互联网络的整个路径被划分为几个部分，每一部分对应于从路由器到路由器，通过特定网络的路径。

为了在下一个网络上传输分组，网络层对应于特定的数据链路技术，将分组放置在帧的数据域里，并且标明下一个路由器接口的数据链路层地址。这个使用适当数据链路层技术的网络，使用指定的地址发送封装着分组的帧。路由器从发送的帧中取得分组，对它进行一些处理，然后将它传送到下一个网络，进行进一步的传输。在做这些之前，它必须将分组封装到新的数据链路层帧中。对应于不同的技术，这一帧可能有不同的格式。因此，网络层基于不同的技术，承担了网络的协同操作的任务。

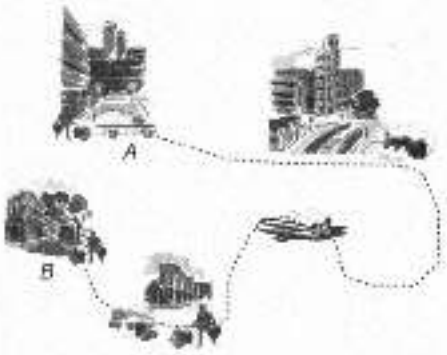


图4-10 国际邮政服务的运作

示例　网络分组的工作方式有些类似于国际邮政服务，例如DHL或者TNT（图4-10）。假设有些货物需要从城市*A*递送到城市*B*，它们在不同的洲。为了递送货物，国际邮政服务可能会用到不同地区的服务提供商，包括：

- 铁路
- 海洋运输
- 航空公司
- 摩托快递

这些运输公司可以被类比为数据链路层网络，每一种网络都建立在特定的技术基础之上。国际邮政服务必须基于这些地区性的服务提供商，组织统一的、协调良好的网络。为了达到这一目标，国际邮政服务必须首先计划货物的运输路径，然后在地区性服务提供商改变的地方协调运作。例如，这可以是从一辆货车上将货物卸下，货物然后被放入飞机的货舱。每一个运输服务提供商负责它自己部分路径的货物的传输，而不负责它自己部分之外货物的运输和状况。

总的来说，网络层的功能不仅仅是在互联网络内保障数据的交换。例如，对于网络内不希望传输的数据，网络层解决了创建可靠、灵活的屏障的问题。

为了总结网络层，我们要指出这一层有两类协议。可**被路由的协议**（Routed protocol）是第一类的协议，它实现了在网络上进行分组转发。这些协议就是我们在讨论网络层协议时经常提到的那些协议。然而，还有另外一种协议——**路由协议**（routing protocol），它们也常常被分类为网络层协议。使用路由协议时，路由器基于分组转发所选择的路由，在互联网络链路的拓扑上收集信息。

4.3.5　运输层

分组在从发送端到接收端的过程中，可能会丢失或者损坏。虽然有些应用程序有它们自己的差错控制工具，但另一些应用程序更喜欢操作可靠的连接。

运输层（transport layer）为应用程序或OSI模型的上层——应用层、表示层、会话层——提供了所需可靠性的数据传输服务。OSI模型定义了五级运输服务，从0级（最低级）到4级（最高级）。根据它们提供的服务质量区分这些级别：紧急；恢复被损坏连接的可能性；使用公共运输协议在不同应用层协议间获得对多个连接进行多路复用的工具的可能性；以及最重要的，检测和纠正传输错误（例如，数据出错，分组丢失，或者重复）的能力。

选择运输层的服务类别一方面取决于应用程序和更高层协议在何种程度上解决了可靠性问题。另一方面，这一选择也取决于网络数据传输系统的可靠性，这由更低层的协议来保障——网络层、数据链路层、物理层。例如，如果通信链路的质量很好，下层协议不能检测出错误的概率很小，那么使用轻量级的运输层服务会是较好的选择，这样的服务不会因为多重检查、确认回执、以及其他提高可靠性的技术而增加负担。如果下层的运输工具不够可靠，那么我们需要使用更高级的运输层服务，它们使用最大可能数量的差错检测和纠正工具，包括逻辑连接建立，根据校验和循环分组编号进行分组传输控制，以及建立传递的超时。

所有运输层及以上层的协议都由安装在网络终端节点的软件工具——网络操作系统的部件实现。运输协议的例子包括TCP/IP协议栈的TCP协议和IP协议，以及Novell协议栈的SPX协议。

七层模型中的四个低层协议被称为**网络运输**（network transport），或者是**运输子系统**（transport subsystem），因为它们基于不同的技术，解决了指定的服务等级，在任何拓扑的网络中进行报文传递的问题。三个高层协议解决了使用运输子系统提供应用服务的问题。

4.3.6 会话层

会话层（session layer）保障了多个参与方之间交互的控制。它登记参与方，提供会话同步的工具。这些工具将检查点插入长时间的传输中，来保证在失败时，回到最新的检查点，而不是从头开始。事实上，使用会话层的应用程序并不多。这一层很少以独立协议体的形式实现。通常，这一层的功能和应用层的功能结合在一起，用单个协议实现。

4.3.7 表示层

表示层（presentation layer）处理通过网络传输的信息的表示形式，而不改变它们的内容。由于这一层的功劳，由一个系统应用层传输的信息总是能被另一个系统的应用层所理解。通过使用这一层的工具，应用层协议可以克服数据表示或不同编码之间的差异，例如，在ASCII和EBCDIC之间的差别。数据加密和解密也由这一层完成，它们保证了所有应用服务数据交换的安全性。这类协议的一个例子是**安全套接字层**（Secure Socket Layer，SSL），它保证了TCP/IP协议栈应用层协议的安全报文交换。

4.3.8 应用层

应用层（application layer）实际上是用户用来访问共享网络资源，例如文件、打印机、网页等的一系列协议。它们也可以通过使用诸如电子邮件协议来组织团队工作。应用层操作的数据单元通常被称为**报文**（message）。

应用层的服务有许多。最著名的网络文件服务实现的例子包括TCP/IP协议栈中的NFS和FTP，微软视窗中的SMB，以及Novel NetWare中的NCP。

4.3.9 OSI模型和电路交换网络

正如以前提到的那样，OSI模型描述了*分组交换网络*（*packet-switched network*）中设备间的交互过程。那么在*电路交换网络*中又如何呢？对于这样的网络是否有参考模型？是否可能将电路交换技术的功能和OSI模型的层进行比较？

当然，电路交换网络的网络互连工具也使用了多层的方法，根据这一方法，各种协议有很多层，共同组成了一个层级结构。但是，对于电路交换网络，没有类似于OSI模型那样的公共参考模型。例如，不同类型的电话网络使用针对特定网络的协议栈，层的数量、以及它们之间不同的功能分布。诸如同步数字层（SDH）或密集波分复用（DWDM）的传输网络也有它们自己的协议层级。因为大多数这种类型的当代网络使用电路交换网络仅仅用来传输用户数据，情况就变得更为复杂了。为了控制连接建立和通用网络管理的过程，人们使用了分组交换技术。例如，SDH、DWDM和其他当代电话网络使用了这一方法。

虽然电路交换网络有相当复杂的组织结构，并且支持它们自己的协议层级，但它们为分组交换网络提供了物理层的服务。

让我们考虑使用数字电话网络将几个分组交换的局域网互联起来的例子。显然，网络互联的功能被赋予了网络层的协议，使得在每个局域网中必须安装一个路由器。这一路由器必须配备有一个接口，这一接口需要能通过电话网络，与另一个局域网建立连接。当这样一个连接被建立起来后，恒定速率的比特流将会被传输到电话网络中。这一连接将为路由器提供物理层的服务。为了组织数据传输，路由器必须在电路上使用某些点对点的数据链路协议。

4.4 网络标准

对大多数技术而言，标准化的好处对计算机网络有特殊的意义。网络的主要思想是保证不同类型的设备之间的通信。因此，兼容性是最重要、最紧迫的问题之一。如果不接受某些设备和协

议的标准，那么网络领域就不会有任何进步。正因为这样，网络的整个演化过程反映在标准里。只有当技术反映在适当的标准中时，它才得到合法化的地位。

之前介绍的OSI模型是计算机网络中标准化的思想体系的基础。

4.4.1 开放系统的概念

什么是开放系统？

在广义上说，**开放系统**（open system）是根据开放标准建立的任何系统（可以是单独的计算机，计算机网络，操作系统，应用程序软件，或是任何其他的硬件或软件）。

在计算中，术语**规范**（specification）意味着：硬件和软件模块的标准化描述、它们运行的方法、它们和其他部件的交互、运行的条件，以及其他的特性。显然，不是每一个规范都是标准。

开放规范（open specification）是发布的、可公开获得的规范。它们经过所有感兴趣的群体间彻底的、繁多的讨论，遵从已达成的标准。

在创建新的系统中，使用开放规范可以使得第三方为系统开发多种硬件或软件扩充，并进行修改。它也允许系统集成商将不同生产厂商提供的硬件和软件产品组合起来。

标准和规范的开放特性不但对于通信协议非常重要，而且对于构建网络所使用的硬件设备和软件产品也非常重要。现在所采用的大多数标准都是开放的。私有系统的精确规范只为它们各自的生产厂商所知道，它们的时代早就结束了。每个人都意识到具有能和竞争产品平稳交互的能力并不会降低产品的价值。恰恰相反，这一价值被极大地提高了，因为这一产品可以使用在基于不同厂商设备所构造的异构网络中。因此，即使对于那些只生产私有系统的公司，例如IBM、Novell、微软，它们现在也积极参与开发开放标准，在它们的产品中实现这些标准。

不幸的是，对于真实世界的系统，完全的开放性依然只是一个不可实现的理想。通常，即使对于号称开放的系统，也只有支持外部接口的特殊部分是真正开放的。例如，操作系统UNIX家族的开放特性包括具有在操作系统内核和应用程序之间的标准软件接口，它使得将一个版本的UNIX系统应用程序移植到另一个版本的环境变得非常容易。

OSI模型和开放性的一个方面相关，即和计算机网络中设备交互的开放性相关。这里，开放系统被解释为通过使用定义好的格式、内容，发送和接收报文意义的标准规则，可以和其他网络设备交互的网络设备。

如果两个网络都根据开放系统的原理构建，就会有以下的优点：

- 构建网络时，可以使用支持同一标准的、不同厂商生产的硬件和软件
- 可以将现有的网络部件简单地替换为更高级的部件，这样就可以以最小的成本扩充网络
- 可以和其他网络方便地互连
- 网络维护的便捷

4.4.2 标准的类型

计算机网络中的所有标准化活动都由多个组织完成的。根据不同组织的性质，标准可以被分为以下类型：

- **私有标准**（Proprietary standard）——这些标准中的一部分是SNA协议栈，它们是IBM的财产，UNIX系统的OPEN LOOK图形接口是Sun微系统公司的财产。
- **专门委员会标准**（Standards of special committee）——这些标准由多个公司共同创立，例如，ATM技术标准由专门建立的ATM论坛开发，这一组织包括一百多家参与公司。再比如，快速以太网联盟为100Mb/s以太网订立标准。
- **国家标准**（National standard）——这些标准包括光纤分布式数据接口（FDDI），它是由美

国国家标准协会（ANSI）开发的众多标准中的一个；操作系统安全标准，它是由美国国防部国家计算机安全中心（NCSC）开发的。

- **国际标准**（International standard）——包括ISO开发的OSI模型和通信协议栈，ITU开发的众多标准，包括X.25、帧中继、ISDN网络和调制解调器的标准。

有些标准在它们的发展过程中可以从一个类别移到另一个类别。例如，流行的、广为使用的产品的私有标准往往变成事实上的国际标准，因为全世界的生产商都必须遵照这些标准，来保证它们产品的兼容性。例如，由于IBM PC的成功，IBM PC体系结构的私有标准成为了事实上的国际标准。

此外，由于某些广为使用的私有标准的流行性，它们成为了国家标准和国际标准的基础。例如，以太网标准最早由Digital Equipment、Intel、Xerox公司开发，过了一段时间，经过一些微小的改动，被采纳为IEEE 802.3国家标准。后来，ISO同意它成为ISO 8802.3国际标准。

4.4.3 因特网标准

因特网（Internet）是开放系统的最好例子。这一网络根据开放系统的要求而发展。成千上万的IT专业人士——全世界不同的大学、科研机构、硬件和软件厂商中这一网络的用户——参与了这一网络标准的开发。定义因特网运作的标准被称为**请求评注**（Request For Comment，RFC）。这一命名强调了标准被采纳的公开性和开放性。因此，因特网成功地将众多散布于全世界各地的不同网络设备和软件聚合起来。

由于因特网广为流行，RFC成为了事实上的国际标准。大多数的RFC后来成为了官方的国际标准，它们通常会得到之前所列出的机构之一的批准（通常是ISO或ITU-T）。

多个组织负责开发因特网的体系结构和协议，特别是对体系结构和协议进行标准化。**国际互联网协会**（Internet Society，ISOC）担任了最重要的职责，它是一个大约有100 000个成员的科学和管理社区。这一社区参与因特网发展的各个方面，并处理社会、政治、技术问题。ISOC协调**因特网体系结构委员会**（Internet Architecture Board，IAB）的工作，后者的职责包括协调TCP/IP栈的研发。这一组织是批准新因特网标准的最高权威机构。

IAB由两个主要的工作组构成：**因特网研究特别任务组**（Internet Research Task Force，IRTF）和**因特网工程师特别任务组**（Internet Engineering Task Force，IETF）。IRTF协调和TCP/IP相关的长期研究项目。第二个工作组，IETF，是一个工程组，负责解决因特网现有的技术问题。IETF定义规范，这些规范最后变为因特网标准。开发和批准因特网标准的过程由七个必备的状态构成。

根据因特网的开放原理，所有的RFC都可以免费访问。所有文档的列表可以在RFC站点找到：http：//www.rfc-editor.org。所有的RFC都可以免费下载。这和ISO的标准不同。

4.4.4 通信协议的标准栈

在计算机网络领域，标准化最重要的方向是通信协议的标准化。最著名的协议栈包括OSI、TCP/IP、IPX/SPX、NetBIOS/SMB、DECnet、SNA，虽然现在并不是上述所有的协议栈都还在实际使用。

1. OSI栈

需要清楚地区分**OSI模型**（OSI model）和**OSI协议栈**（OSI protocol stack）的概念。OSI模型是在开放系统之间进行交互的概念性方法，与之不同的是，OSI栈是一系列特定协议的规范。

与其他协议栈不同的是，OSI栈（图4-11）完全符合OSI模型，并包括了这一模型定义的七层交互的协议规范。这并不意外，这一栈的开发者使用OSI模型作为参考和实际开发的指南。

OSI栈的协议具有复杂和规范模糊的特点。它们的特性反映出栈协议开发者的共性，他们试图考虑所有已经存在和将要出现的技术。

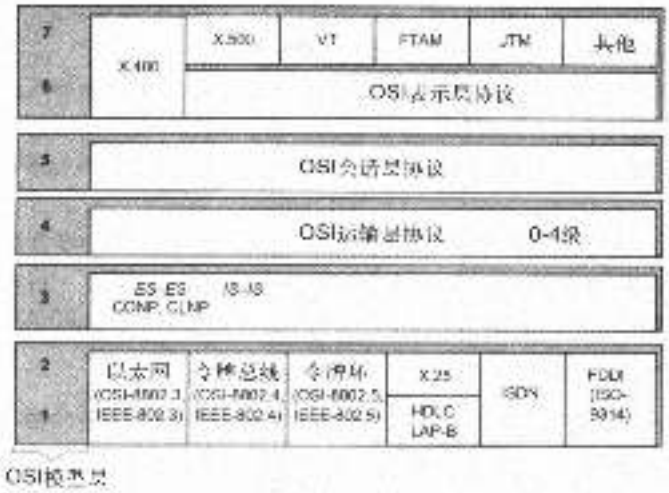


图4-11　OSI协议栈

在物理层和数据链路层，OSI栈支持诸如以太网、令牌环、光纤分布式数据接口、LLC、X.25、ISDN这样的协议。换句话说，它使用了在栈的框架外开发的低层次协议。在这里，它和大多数其他的协议栈类似。

网络层、运输层、会话层的服务也出现在OSI栈中，虽然它们实际上很少被使用。在网络层，既可以使用面向连接的网络协议（CONP），也可以使用无连接的网络协议（CLNP）。使用这些协议，以下两个路由协议也会被用到：终端系统-中间系统（ES-IS）和中间系统-中间系统（IS-IS）。

根据OSI模型中为运输层协议定义的功能，OSI栈的运输层协议隐藏了面向连接的服务和无连接服务之间的区别，以致于用户独立于下面的网络层得到所要求的服务质量。为了保证这一点，运输层要求用户指定所需要的服务质量。

应用层包含文件传输、终端仿真、目录服务以及电子邮件。最常见的服务是目录服务（X.500标准）、电子邮件（X.400）、虚拟终端协议（VTP）、文件传输和访问，管理协议（FTAM）、工作传送与管理协议（JTM）。

2. IPX/SPX栈

IPX/SPX栈（the IPX/SPX stack）最早是由Novell公司在20世纪80年代早期为它自己的网络操作系统NetWare开发的协议栈。图4-12阐述了IPX/SPX栈的结构，以及它和OSI模型的对应关系。网络层和运输层的协议——互联网分组交换（IPX）和顺序分组交换（SPX）——用它们的名字命名了整个协议栈。路由协议RIP和NLSP也被关联到了这个栈的网络层。最高三层的代表是远程访问NetWare文件的协议——NetWare核心协议（NCP）和服务广告协议（SAP）。

> **注意**　直到1996年，就安装数量而言，该一协议还是毫无争议的世界冠军，但是，形势突然改变了，TCP/IP栈开始各项长率和安装数量上超过了它的栈，到1998年，TCP/IP成了绝对的领导者。

许多IPX/SPX栈的特定特点可以归因于早期NetWare版本对由PC机组成的小型局域网的定位，这些PC机具有有限的资源。为了实现这一目标，Novell要求协议的实现只需要使用最小数量的随机存储器（随机存储器的容量在运行MS-DOS的IBM兼容PC机上相当有限——只有640KB）。此外，

这些协议必须保证能高速运行在这些功能较弱的机器上。正是因为这样，IPX/SPX栈的协议至今还在局域网中工作良好。但是，在大规模的公司网络中，由于这一栈中某些协议大量使用广播分组的方法（例如，SAP），它们能够慢慢的全球链路产生超负载。此外，IPX/SPX栈是Novell公司的财产，为了实现这一栈，必须购买一个许可证，这意味着它并不支持开放规范。这些状况长期限制了它的使用，使它仅仅用于NetWare网络中。

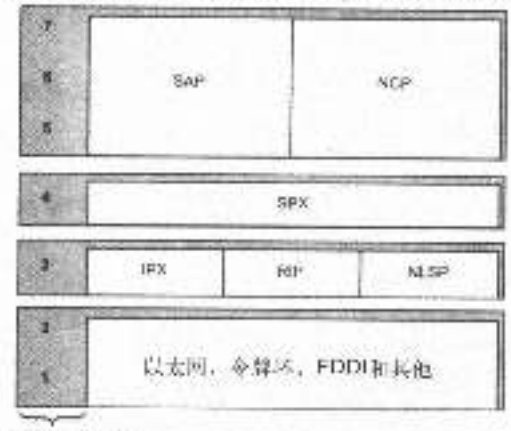


图4-12 IPX/SPX协议栈

3. NetBIOS/SMB栈

NetBIOS/SMB栈（NetBIOS/SMB stack）由IBM和微软开发（图4-13）。这一栈的物理层和数据链路层使用了大部分流行的协议，包括以太网、令牌环、光纤分布式数字接口。这一栈上面的几层使用了NetBEUI和SMB协议。

网络基本输入/输出系统（NetBIOS）协议作为IBM PC网络软件的标准IBM PC BOIS功能的网络扩充，首先出现在1984年。后来，这一协议由NetBIOS扩展用户接口（NetBEUI）协议替代，为了保证应用程序的兼容性，NetBIOS接口作为NetBEUI协议的接口被保留下来。NetBEUI被开发为一个对计算机资源只有低要求的高效的协议，它适用于由小于200台工作站构成的小型网络。这一协议实现了大量有用的网络功能，它们位于OSI模型的运输层和会话层。不幸的是，这一协议不允许分组路由。这将NetBEUI协议的使用限制在没有划分子网的小型局域网中，使它无法在互连网络中使用。

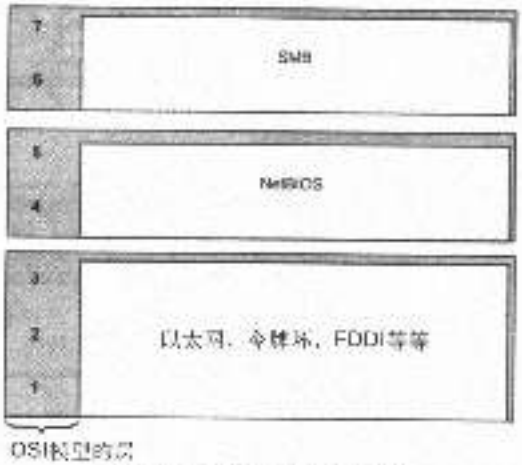


图4-13 NetBIOS/SMB协议栈

服务器报文块（SMB）协议实现了OSI模型中会话层、表示层、应用程的功能。SMB是实现文件服务、打印和消息服务的基础。

4. TCP/IP栈

TCP/IP栈（TCP/IP stack）由美国国防部在二十多年前发起开发，它的目的是为了保证实验性的ARPANET网和其他网络的互连。TCP/IP被实现为对异构网络环境的协议组。加州大学伯克力分校为TCP/IP栈的开发做出了最大的贡献，他们在流行版本的UNIX操作系统中实现了TCP和IP协议，TCP/IP栈就用这一栈中最著名的这两个协议命名的。UNIX的流行使得TCP和IP，以及这一栈中其他的协议得到了普遍的应用。现在，这个栈用于连接到因特网的计算机之间的通信，也用于大量的公司网络中。

因为TCP/IP栈最早为因特网开发，和其他协议相比，它有许多优点，特别是在构建具有广域网链路网络的时候。特别地，TCP/IP将分组分段的能力非常有用，使得在大规模网络中运用这一协议栈成为可能。大规模互联网络通常建立在各种不同的网络之上，这些网络基于不同的原理，对于每一个这样的网络，可能会有一个和网络相关的最大传输单元的值（帧的大小）。这时，当数据从具有大的最大帧长的网络传输到具有小的最大帧长的网络时，可能需要将传输的帧分为几个

帧。TCP/IP栈的因特网协议有效地解决了这一问题。

灵活的编址系统是TCP/IP技术的另一大优点。这一性质也鼓励了在大规模、异构网络中使用TCP/IP。

TCP/IP栈很少使用它的广播能力。在使用慢速链路的时候，这一性质绝对必要。而慢速链路在长距离网络中经常使用。

但是，和以前一样，优点总是会有一些代价。对它而言，之前列出的优点是以对资源的高要求和复杂的IP网络管理为代价的。TCP/IP栈强有力的功能需要相当多的资源。灵活的编址和放弃广播使得在IP网络中需要多种集中服务，例如域名系统（DNS）和DHCP。每一种服务都是为了协助网络管理的过程。然而，每一种服务都需要网络管理员予以密切的关注。

我们也可以给出TCP/IP栈的其他优缺点。但无论如何，这一协议栈是当今最流行、最广为使用的协议栈。不仅在广域网中是这样，在局域网中也是这样。

图4-14介绍了TCP/IP栈的体系结构。和OSI模型类似，TCP/IP栈也有多层结构。然而，TCP/IP栈的开发早于ISO/OSI模型，TCP/IP栈的层并不完全对应于OSI模型的层。

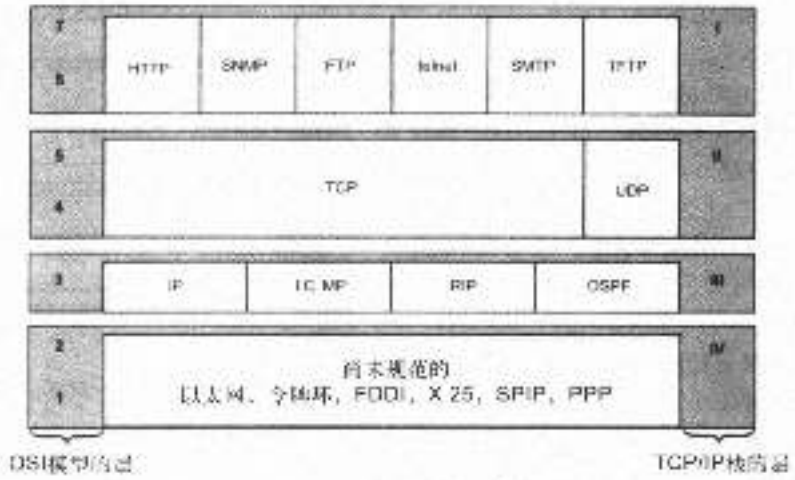


图4-14　TCP/IP栈的体系结构

TCP/IP栈定义了四层：

TCP/IP栈的应用层（*application layer*）对应于OSI模型的上三层：应用层、表示层、会话层。它将系统提供的服务和用户应用结合起来。在不同国家、不同机构的不同网络的运作过程中，TCP/IP栈在应用层积累了大量的协议和服务。这些协议和服务的列表相当之长，包括广为使用的协议，如文件传输协议（FTP）、终端仿真协议（telnet）、简单邮件传输协议（SMTP）和超文本传输协议（HTTP）。

应用层协议安装在主机（host）上[⊖]。

TCP/IP栈的运输层（*transport layer*）可以为上一层提供如下两类服务：

- 保证递送　传输控制协议（TCP）
- 尽力递送——用户数据报协议（UDP）

为了保证可靠的数据传输，TCP提供了建立逻辑线路的方法。这使得它能给分组编号、确认回执、重新传输任何丢失的分组。它还可以检测并丢弃重复的分组，以分组发送的顺序递送分组到

⊖ 在因特网术语中，终端节点传统上称为主机（host），路由器称为网关（gateway）。在本章中，我们也使用这些术语。

应用层。这一协议允许源计算机和目标计算机上的对象支持双工模式的数据交换。TCP保障了一台计算机上形成的字节流能无错地传输到连在互联网络上的另一台计算机。TCP将字节流分割为小段，将它们传递到下面的网络层，在网络互连工具将它们传送到目的计算机后，TCP再将它们组合成一个连续的字节流。

这一层的第二个协议——UDP——是一种最简单的数据报协议，当不存在可靠数据交换的问题，或是这一问题已由更高层，例如应用层，或是用户程序解决的情况下，它往往会被使用。

TCP和UDP的功能包括了在应用层和下面的互联网层之间的链路功能。运输层根据应用层所指定的质量，担任数据传输任务，在完成任务后，在再通知应用层。另一方面，TCP和UDP使用将下面的网络层作为一种工具，网络层不具有高可靠性的特点，但可以通过互联网络传送分组。与应用层协议类似，TCP和UDP安装在主机上。

网络层（*network layer*）也被称为互联网层（*internet layer*），它是整个TCP/IP体系结构的核心。这一层的功能对应于OSI模型中的网络层，它保证分组在互联网络中转发，互联网络由多个网络连接而成。网络层的协议支持对上面运输层的接口，从运输层接收数据传输的请求，网络层还支持对下面的网络接口层的接口，网络接口层的功能将在后面介绍。

因特网协议（IP）是网络层的主要协议。它的任务包括在网络间进行分组转发——从一个路由器到另一个路由器，直到分组到达目的网络。和应用层和运输层协议不同，IP不但安装在所有的主机上，还安装在所有的网关上。IP是无连接数据报协议，根据尽力原则运作。

完成IP协议辅助功能的协议也常被划分到TCP/IP网络层。这些协议包括诸如路由信息协议（RIP）和开放最短路径优先（OSPF）协议等路由协议，它们学习网络拓扑、决定路由、创建路由表，帮助IP将分组向需要的方向转发。出于同样的原因，另外两个协议也可以被划分到网络层：因特网控制报文协议（ICMP），它从路由器向信息源传送分组传输错误信息；因特网组管理协议（IGMP），它用来同时向多个地址转发分组。

TCP/IP栈的体系结构和其他多层结构协议栈的区别在于其对最低一层功能的不同理解。这些功能就是网络接口层（*network interface layer*）的功能。

让我们回忆一下，OSI模型的最低层（数据链路层和物理层）实现了许多负责媒体访问的功能：帧的操作、协调不同电信号的水平、编码、同步等等。所有这些相当专业的功能是诸如以太网、令牌环、PPP、HDLC、以及其他许多数据交换协议最重要的部分。

TCP/IP栈的低层解决了一个更为简单的问题：它只负责组织和其他网络技术的交互，这些技术用在构成互联网络的网络中。TCP/IP将互联网络中的任何网络作为整个路由中将一个分组传输到下一个路由器的工具。

因此，在TCP/IP技术和其他中间网络技术之间提供接口的任务被简化为如下任务：

- 定义将IP分组封装为中间网络PDU的方法
- 决定将网络地址转化为中间网络技术使用的地址的方法

这样一种方式使得TCP/IP互联网络非常开放，其他任何网络都能够加入其中，无论它们内部使用何种数据传输技术。对于互联网络中各个网络使用的各种技术，我们需要开发特定的接口工具。因此，这一层的功能不能一劳永逸地定义出来。

不能严格地规定TCP/IP栈中的网络接口层。它支持所有流行的网络技术。对于局域网，它们是以太网、令牌环、光纤分布式数字接口、快速以太网、千兆以太网；对于广域网，它们是诸如SLIP和PPP的点对点协议；对电路交换网络，它们是X.25、帧中继、ATM技术。

通常，当任何新的局域网或广域网技术出现时，通过开发合适的RFC，确定封装分组到帧的方式，这些新技术就能迅速地包含到TCP/IP栈中。例如，RFC1577定义了ATM网上的IP操作，它

出现于1994年，就在它被采纳为ATM的主要标准后不久。

注意　TCP/IP栈允许将网络包含进互联网络中，而无论这些网络的层数有多少。因此，举个例子说，X.25网络中的数据转发由物理层、数据链路层、网络层（OSI的术语）的协议保障。然而，TCP/IP栈仅仅将X.25网络和其他的技术认为是在两个边界路由器间传输IP分组的工具。网络接口层提供了将IP分组封装成X.25分组，以及将IP地址转化为X.25网络层地址的方法。如果这一网络的结构和OSI模型完全吻合，我们就要承认一个明显的矛盾——一种网络层协议（IP）工作在另一种网络层协议之上（X.25），然而，对于TCP/IP栈来说，这是很正常的。

每一种通信协议操作特定的所传输的数据单元。这些数据单元的命名规范有时候在标准里指定；更为常见的，它们根据传统习惯来决定。在TCP/IP栈的长期存在过程中，特定术语经常被使用（图4-15）。

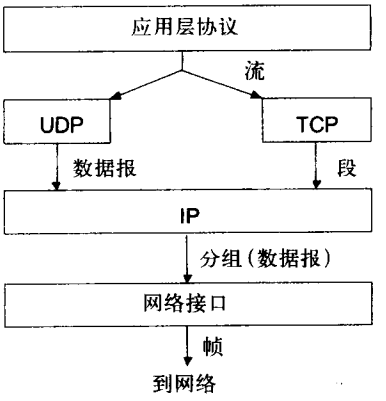


图4-15　TCP/IP PDU名称

流（stream）是用来指代从应用程序到运输层协议（TCP、IP）输入处的数据的术语。

TCP将流划分为若干**段**（segment）。

UDP的协议数据单元常被称为**数据报**（datagram）。数据报是无连接协议中使用的PDU的常见名称。这类协议包括IP，因此，它的PDU也被称为数据报。但是，另一个术语也很常用：IP分组。

根据TCP/IP栈的术语，**帧**（frame）是IP分组为了之后在网络上的传输，而被封装进去的任何技术的PDU，这些网络是互联网络的一部分。这时，在特定网络技术中，这一PDU使用何种名字并不重要。因此，以太网帧、ATM信元、X.25分组都被TCP/IP栈认为是帧，因为所有这些PDU都被认为是容器，它们容纳IP分组，在互联网络上传输。

4.4.5　流行协议栈与OSI模型间的对应关系

图4-16显示了标准协议栈和OSI参考模型之间的对应关系。正如你看到的那样，这种对应关系相当一般。在大多数情况下，栈的开发者牺牲模块结构来加速网络运行。只有OSI栈被分为七层。更为常见的是，栈被分为三到四层：网络适配器层，它实现了OSI模型的物理层和数据链路层；网络层；运输层；服务层，它将OSI模型中会话层、表示层和应用层的功能结合在一起。

除此之外，协议栈的结构和OSI建议不一致还有其他的原因。让我们回忆一下理想的多层分解的特性。首先，需要遵守层次性的原理：任何上层都只能请求它下面的那一层，任何下层都只能给它上面的那一层提供服务。在协议栈中，这将造成上层PDU总是被封装进下层PDU中。

其次，理想的多层分解假设所有同一层的模块负责完成一个共同的任务。然而，这一要求常常和实际的协议栈的构造相矛盾。例如，TCP/IP栈网络层的主要功能（与OSI栈网络层类似）是保障通过互联网络传输分组。TCP/IP栈提供多个协议解决这一问题，例如IP，它用来转发分组，以及RIP和OSPF，它们是路由协议。假如协议解决共同的问题，这些协议就应该属于同一个层，那么，因特网和路由协议应该属于同一层。然而，RIP报文被封装进UDP数据报，OSPF报文被封装进IP分组，因此，如果严格遵守栈的层次结构，OSPF协议应该属于运输层，而RIP应该属于应用层。但是，事实上，路由协议通常为认为是网络层的协议。

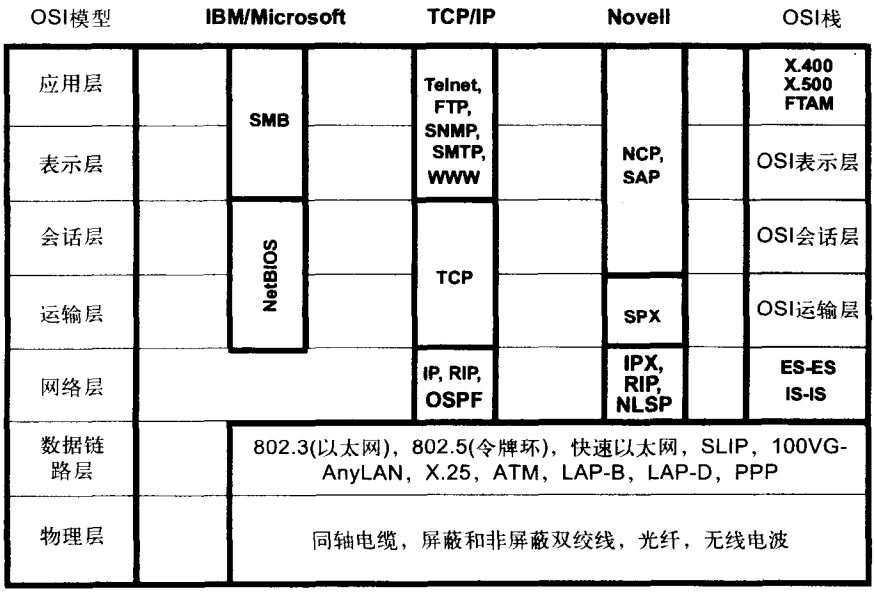


图4-16 流行的协议栈和OSI模型各层之间的对应关系

4.5 信息和运输服务

计算机网络的服务可以被分为以下两类：

- 运输服务
- 信息服务

运输服务（transport service）假设数据以不变的形式在网络用户之间传送。网络在它的接口之一输入用户数据，通过中间交换机传送数据，通过另一个接口将数据输出给另一个用户。在提供运输服务时，网络并不改变传送的信息。它以发送端提供给网络的同样的形式，将数据传送到接收端。广域网提供的一个运输服务的例子是客户端局域网的网络互连。

信息服务（information service）包括提供新信息给用户。信息服务总是和数据处理操作相关：将数据以某种有序的方式（文件系统或数据库）存储，搜索需要的信息，将它以所需要的形式呈现出来。信息服务的出现远远早于第一个计算机网络。电话网络提供的电话目录查找服务是信息服务的一个典型例子。随着计算机的出现，由于计算机是为了自动信息处理而发明，信息服务经历了一次革命。不同的信息技术，包括编程、数据库、文件存储、万维网、电子邮件，都被用来提供信息服务。

在计算机时代前的电信网络中，运输服务非常盛行。电话网络的主要服务就是在两个用户之间传输语音流量。查询服务处于辅助的地位。在计算机网络中，这两种服务同样重要。计算机网络的这一特点反映在新一代电信网络的名字上，后者是将不同类型的网络融合在一起的产物。现在，这样的网络常常被称为**信息通信网络**（infocommunication network）。虽然这一名称还没有被普遍地接受，但它很好地反映出新的趋势，将网络服务的两个部分置于同一层次上。

将计算机网络服务划分成两类已经在很多地方显示出来了。例如，现在在计算机网络领域，专业工作者被严格地分类。他们是IT专业工作者和网络专业工作者。第一类的专家包括程序员、数据库开发者、操作系统管理员、网站设计师——也就是说，所有参与计算机软件和硬件开发与支持的专业工作者。第二类包括参与解决网络运输问题的专业工作者。他们与通信链路和通信设备打交道，例如交换机、路由器、集线器。他们解决选择网络拓扑、定义通信流的路由、定义链路

的带宽需求和通信设备的性能，以及其他使用网络传输数据的问题。

毫无疑问，每一类专业工作者都需要了解相邻领域的问题和方法。参与开发分布式应用的专业工作者必须知道他们可以从网络得到何种运输服务，以便于在分布式的应用部件中进行协调操作。例如，网络程序员必须知道TCP/IP栈提供的两种不同的运输服务——TCP和UDP。因此，程序员将决定哪种服务更适合于特定的应用。类似地，网络运输工具开发者必须理解应用程序进行数据传输的需求，并在网络设计时考虑这些需求。尽管如此，IT领域和网络领域的专业性仍然存在，这反映出计算机网络的双重目的。

将网络服务划分为运输服务和信息服务也反映在协议栈的组织上，并且，还反映在各种协议栈部件的分布上。

4.5.1　网络元素的协议分布

图4-17显示了计算机网络的主要部件：终端节点，或计算机，以及中间节点，或交换机和路由器。这个例子选择了TCP/IP栈的协议，因为它们是最常用的。

从图上我们可以很明显地看出，完整的协议栈仅仅在终端节点上实现，中间节点支持最底下三层的协议。最底下三层协议的功能有助于解释这一原因——它们对于分组转发的通信设备已经足够了。此外，通信设备可以仅仅支持最底下的两层协议，甚至只支持物理层——这取决于设备的类型。

- **集中器**（Concentrator）——它对比特流操作，因此，它被限制在对物理层协议的支持。
- **局域网交换机**（LAN switche）——它们支持最低的两层协议——物理层协议和数据链路层协议。因此，它们可以在标准拓扑结构的范围内操作。
- **路由器**（Router）——由于它们需要网络层，所以互连基于不同技术的网络，因此，它们必须支持最下面的三层协议。最下面的两层协议用于和组成互联网络的成份网络交互（例如，以太网或帧中继）。
- **广域网交换机**（WAN switche）——广域网的交换机，例如ATM，运行在可以支持两层或二层协议的虚电路的基础上。支持自动建立虚电路所需要的网络层协议。由于广域网的拓扑结构是任意的，所以没有网络协议就不行。另一方面，假如虚电路由网络管理员人工设立，为了在已有的虚电路上传输数据，广域网交换机只支持物理层和数据链路层协议也就够了。

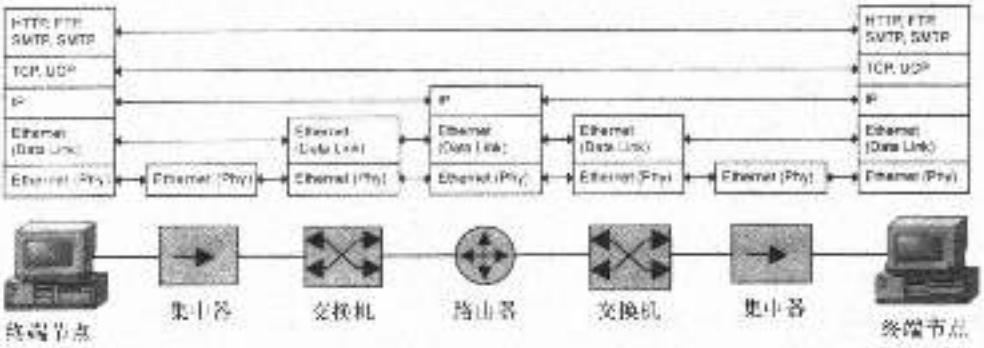


图4-17　网络设备的功能和OSI模型层的对应关系

运行网络应用程序的计算机必须支持所有层的协议。应用层协议使用表示层和会话层提供的服务，以网络API的形式向应用程序提供一套网络服务。运输层协议也运行在所有终端节点上。当需要使用网络组织数据传输时，两个运行在发送节点和接收节点的运输协议实体互相交互，以保障提供运输服务所需要的质量。网络通信设备透明地递送运输协议的报文，而不需要知道它们的内容。

在计算机中，所有层的通信协议（除了物理层和数据链路层的某些功能）由系统软件实现。

网络的终端节点（计算机和基于计算机的高科技设备，例如移动电话）总是提供信息服务和运输服务，中间网络节点只提供运输服务。如果某个网络只提供运输服务，那意味着终端节点在网络的边界之外。通常，这是提供服务给客户的商业网络的情况。如果我们说，一个网络也提供信息服务，这意味着提供这些服务的计算机包括在网络中。一个典型的例子是在某个网络中，因特网服务提供商除了提供因特网访问服务外，还支持它自己的网站服务器。

4.5.2 运输系统的辅助协议

显然，图4-17描述了经过简化的、在网络元素间的协议分布状况。在现实世界的网络中，有些通信设备不但支持最底下三层的协议，还支持一些高层的协议。例如，实现了路由协议的路由器允许自动建立路由表。集中器和交换机常常支持SNMP和telnet协议，对于这些设备实现的主要功能，这些协议并非是必要的，但提供了远程配置和控制这些设备的方法。所有这些协议都是应用层协议，它们实现了运输系统的一些辅助功能。显然，为了支持应用层协议，网络设备必须支持中间层协议，例如IP和TCP/UDP。

辅助协议可以根据它们的功能，分为不同的组：

第一组包含**路由协议**（routing protocol），例如RIP、OSPF和BGP。如果没有这些协议，路由器不能路由分组，因为路由表将是空的（除非网络管理员人工填入——对于大型网络来说，这并不是一个合适的解决方案）。如果我们不仅仅考虑TCP/IP栈，还考虑支持虚电路网络的协议栈，这一组还将包括用来建立虚电路的协议。

另一组辅助协议参与**地址转换**（address translation）。特别地，它包括了DNS协议，这一协议将字符节点名转换为IP地址。这一组的另一个协议，DHCP，允许为网络主机动态分配IP地址。这不同于静态地址，对于后者，这一地址必须由网络管理员人工指定。因此，网络管理员的工作也被大大简化了。

第三组包括用来进行**网络管理**（network management）的协议。出于这一目的，TCP/IP栈包括了简单网络管理协议（SNMP），它可以自动收集错误信息和设备失败的信息。TCP/IP协议栈还包括了telnet协议，网络管理员可以使用它远程配置交换机和路由器。

在考虑辅助协议时，我们遇到了OSI模型采用的协议分层方式（即，垂直划分）不够用的情形。除了不同层次之外，还需要将垂直划分的协议分为几个组。

虽然OSI模型并没有提供这样一种划分，但事实上存在。这一方法在ISDN网络的标准化过程中用到，正如我们介绍过的那样，ISDN网络同时使用分组交换技术和电路交换技术。ISDN标准将所有的协议分为三个组，称为用户平面（User Plane）、控制平面（Control Plane）和管理平面（Management Plane）（图4-18）。

- **用户平面**（User Plane）组包括用来传输用户语音流量的协议。
- **控制平面**（Control Plane）组包括用来建立网络连接的协议。
- **管理平面**（Management Plane）组将支持网络管理运作的协议（例如差错检测、分析设备配置）连接起来。

从这一描述中，我们可以清楚地发现，在平面功能和基于TCP/IP或其他技术的计算机网络辅助功能组之间，存在着很大的相似性。虽然这种水平的协议划分并没有在计算机网络中被广泛地接受，但它还是非常有用的，因为它帮助人们更好地理解每个协议的目的。除此之外，将某些协议和OSI模型层之间关联起来具有一定的困难，水平划分有助于对此进行解释。例如，某些作者将路由协议放在网络层，而有些作者将它们划入应用层。这并不是因为作者的疏忽大意，而是因为划分起来确实具有一些客观的困难。OSI模型非常适合运输用户流量的协议的标准化（即，可以被分为用户平面的协议）。但是，它并不太适用于辅助协议的标准化。因此，许多作者将路由协议放在网络层，来反映它们和由IP实现的网络运输服务在功能上的相似性。

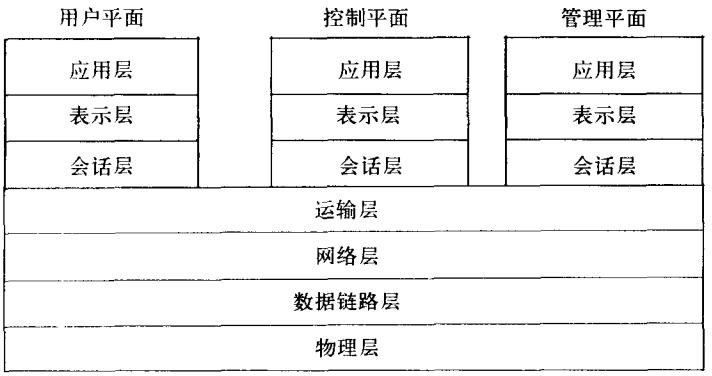


图4-18　三组协议

小结

- 计算机之间交互的高效的网络模式是多层结构，在这一结构中，上一层的模块使用下一层提供的工具来完成它们的任务。每一层支持两类接口——与同一网络节点上层和下层之间的服务接口，以及和远程节点同一层工具的对等接口。后一种接口也被称为协议。
- 协议栈是用来在网络节点间层次性地组织交互的一系列协议。低层的协议通常由软件和硬件的组合来实现。高层协议通常基于软件工具来实现。实现一个特定协议的软件模块称为协议体。
- 在20世纪80年代早期，ISO、ITU-T以及其他在标准化领域工作的国际组织开发了标准OSI模型。这一模型描述了一般化的网络互联工具。网络专业人员用它作为一种通用语言，因此，它被称为参考模型。OSI模型定义了交互的七层结构，给了它们标准的名字，并指定了每一层需要完成的功能。
- 开放系统是根据公开规范建立，符合标准，并且经过所有感兴趣的参与方的公开讨论，而被采用的任何系统（计算机、计算机网络、软件产品、操作系统或任何其他软件或硬件）。
- 根据不同组织机构的状况，可以将它们的标准分为以下几类：特定公司的私有标准、专门委员会开发的标准、国家标准和国际标准。
- 在计算机网络领域，标准化最重要的方向是通信协议的标准化。标准化的协议栈包括：TCP/IP、IPX/SPX、NetBIOS/SMB、OSI、DECnet和SNA。TCP/IP栈具有领导地位，它用来在数千万参与全球信息网络（即因特网）的计算机中进行通信。TCP/IP栈有四层：应用层、运输层、网络层、网络接口层。TCP/IP和OSI层之间并不完全对应。

复习题

1. OSI模型标准化了什么？
2. 是否可能构造一个具有不同层数的OSI模型的变种，例如，五层或八层？
3. 协议是解决系统之间交互问题的软件模块，还是交互规则的标准化描述，包括交换的报文序列和它们的格式？
4. 术语接口（*interface*）和协议（*protocol*）是否是同义词？
5. 应用程序运行在OSI模型的哪一层上？
6. 运输层协议安装在哪些网络元素上？
7. 网络服务运行在OSI模型的哪一层上？
8. 以下哪些设备实现了OSI模型的物理层功能？哪些实现了数据链路层的功能？

a. 路由器
b. 交换机
c. 网桥
d. 转发器
e. 网络适配器

9. 每一层的协议数据单元传统上使用哪些名字？请填在下表里。

	分组（packet）	报文（message）	帧（frame）	流（flow）	段（segment）
数据链路层					
网络层					
运输层					
会话层					
表示层					
应用层					

10. 请给出开放系统的例子。
11. 假设一个不知名的小公司提供给你一个你所需要的产品，它的参数超过了知名大公司提供的类似产品的参数。你可以：检查制造商提供的文件，并确认标出的参数确实超出著名产品的类似参数，然后接受这一产品；或者，经过仔细的测试，确认这一产品的技术参数确实好于市场上的类似产品，然后接受这一产品；或者，选择一家世界知名公司的产品，因为它们的产品一定可以遵守标准，公司没有倒闭的危险，因此，技术支持得到保障。根据开放系统的原理，你会采取何种行动？
12. 哪一组织开发了以太网标准？
13. 因特网的哪些管理组织直接参与标准化？
14. 术语标准（standard）、规范（specification）和RFC是否是同义词？
15. 当代的RFC是哪一类的标准？
    a. 私有标准
    b. 政府标准
    c. 国家标准
    d. 国际标准
16. 哪一个组织发起了TCP/IP栈的创立和开发？
17. 请描述TCP/IP栈的主要性质。
18. 请比较TCP/IP和OSI参考模型最底下几层的功能？
19. 请定义运输服务和信息服务。
20. 哪些协议属于控制平面？哪些属于管理平面？
21. 路由器是否需要支持运输层协议？

练习题

1. 假设通过以太网适配器，你有两台电脑连接在网络上。安装在这些电脑上的适配器驱动程序支持到IP网络层协议的不同接口。这两台电脑是否可以正常交互？
2. 如果两台电脑使用以下各层的不同协议，你如何组织这两台电脑的交互？
   - 物理层和数据链路层
   - 网络层
   - 应用层
3. 假如你需要检查标准化MPLS技术的流程的状态，你需要采取哪些步骤来完成这一任务？
4. 请找出IETF非常关注的领域（例如，通过工作组的数量）。

第5章 网络的例子

5.1 引言

这章描述最流行的几种网络的体系结构——电信运营商的网络，企业范围的网络和因特网。

尽管这几类网络之间存在着差别，它们也有很多共同点。首先，它们具有类似的体系结构。例如，任何电信网络都由主干、接入网、信息中心、以及客户设备构成。自然，针对每种特定类型的网络，这一通用设计有各自特定的信息内容。

电信运营商网络的不同之处在于它们提供了公共的服务。传统意义上，它们提供的服务包括电话服务和租赁线路服务。这样，各种机构可以使用租借的线路建立它们自己的网络。随着计算机网络的流行，电信运营商极大地扩充了他们的服务范围。例如，现在大多数运营商提供因特网访问、虚拟专用网络（VPN）、网站托管、电子邮件、IP电话、语音以及视频的广播。

20世纪80年代中期，这一领域的反垄断开始在全世界范围内出现。其结果是，传统的电信运营商被剥夺了提供公众服务的垄断权。这一过程使竞争运营商出现，它们试图通过使用一套扩充的服务、在服务中达到更好的性价比来吸引客户。理解电信界的管理结构，有助于理解网络技术的一些特定特征，在某些情况下，它们是为特定类型的运营商专门设计的。

企业范围的网络有着和电信运营商网络类似的层次结构，但是，它们通常只向公司的雇员提供服务。因此也有一些不同。

这章的最后部分介绍了因特网。这一网络在很多方面非常独特。它对全世界网络技术的发展有着非常重要的影响力。

5.2 电信网络的一般结构

尽管在计算机网络、电话网络、电视网络、广播网络和电信网络之间存在着不同，但它们的结构中也有一些共同的特征。总体而言，任何电信网络都由以下部分构成（图5-1）：

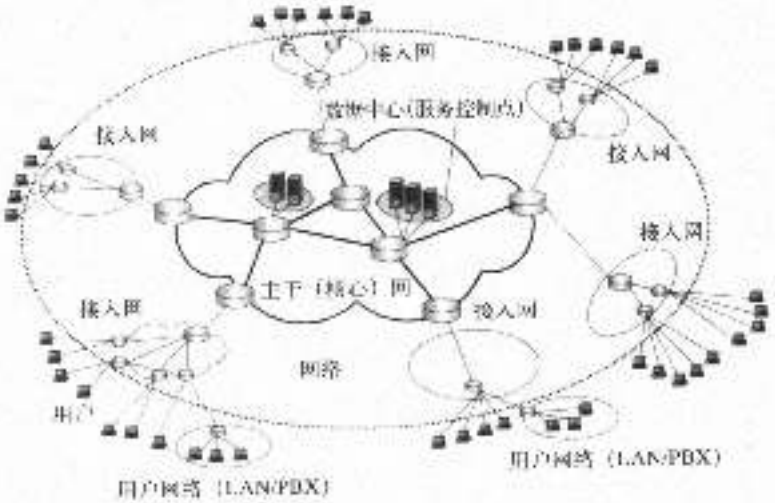


图5-1 电信网络的一般结构

- 数据终端设备（可能连接在网络上）
- 接入网
- 主干（或核心）网
- 数据中心或服务控制点

接入网和主干网都建立在交换机的基础上。每一个交换机都有多个端口，并通过**通信链路**（communications links）和其他交换机的端口连接在一起。

5.2.1 接入网

接入网（access network）构成了电信网络较低的层次。接入网的主要目的是在较少数量的主干网节点中，汇集来自用户设备的、通过多种类型链路到达的信息流。

在计算机网络中，终端节点是计算机；在电话网络中，终端节点是电话机；在电视和广播网络中，终端节点是电视和广播接收器。由于用户的终端设备是终端用户的财产，并且位于用户驻地，这些设备可以被加入到并不属于电信网络的网络中。终端用户的计算机被连接在一起，构成局域网；电话可以被连接到**用户交换机**（Private Branch Exchange）上。

接入网是具有许多分支的区域性网络。与其他电信网络类似，接入网可以由多个层构成。图5-1显示了其中的两个。安装在低层节点的交换机对多个用户信道到达的信息进行多路复用，它们常被称为本地回路，然后交换机将这些信息传输到高层交换机，高层交换机再传输到主干交换机。

接入网的层数取决于它的规模。小的接入网可以只有一层；大的接入网通常包含两层甚至三层。

5.2.2 主干

主干（核心）（backbone/core）网将若干接入网连接起来，通过快速链路，在它们之间传输中间流量。

主干交换机不但可以对单个用户之间的信息连接进行操作，还可以对汇集大量用户数据的信息流进行操作。于是，通过主干网传输的信息到达接收端的接入网，在那里，它们被解多路复用和交换，这样，每个用户的输入端口只接收到要发送给他的信息。

示例 你可以很容易地发现任何国家公路系统具有和大型电信网络同样的层次结构。通常，村庄和小镇由扩展的、局域性的本地道路系统连接起来。这些道路常常很窄；这些定居点之间的流量通常很低，所有没有理由要把这些道路建设成多车道。这些道路连接在省道上，后者往往更宽，并且保障更快的通行速度。省道连接在国道上。这反映了定居点间和国家区域间的交通流量，使得交通运输更有效率。

5.2.3 数据中心

信息（information）或**数据中心**（data center），也可以称为**服务控制点**（service control points，SCP），提供了网络的信息服务。这样的中心可以存储两种类型的信息：

- 用户信息。这是网络终端用户直接感兴趣的部分
- 辅助信息。它们协助提供服务给终端用户

门户网站是信息资源的一个例子，这些网站包括参考信息、新闻、网上商场信息等。在电话网络中，这类中心提供多种服务，诸如紧急呼叫（例如，报警或救护服务），以及对一些机构的查询服务，例如火车站、机场、商店等。

存储第二类资源的信息中心包括：进行用户认证和授权的各种系统；在商业网络中计算服务费用的计费系统；存储用户登录名、密码，以及每个用户所定制的服务的用户账户数据库。电话

网络中有集中式的SCP，在SCP中的每台计算机运行程序对非标准的用户电话进行处理，例如免费商业服务查询（1-800服务），或是参与电话投票的电话。

很自然地，每种类型的网络都有许多特定的特点，但是，它的结构都与之前所描述的结构类似。同时，根据网络的目的和规模，一般结构中的某些部分可能有，也可能没有。例如，小型的局域网没有明显的接入网或干干网，因为它们被合并成为一个共同的、相对简单的结构。通常，企业范围的网络没有计费系统，因为公司以非盈利的原则为雇员提供服务。数据中心可能缺少某些电话网络。在电视网络中，接入网看上去更像一个分布式网络，因为这类网络中的信息流是单向的——从网络到用户。

5.3　电信运营商网络

我们之前提到过，网络分类最重要的特点是其所提供服务的用户的多样性。电信运营商（服务提供商）的网络提供公共服务；企业范围的网络通常只服务于本企业的雇员。

电信运营商（telecommunications carrier）指创建电信网络以提供公共服务，拥有这一网络，并维护它运营的专门公司。

电信运营商根据服务条款，向他们的客户提供商业服务。

各电信运营商在以下方面存在区别：

- 提供的服务
- 提供服务所覆盖的地理范围
- 服务所定位的客户类型
- 运营商拥有的网络基础结构——通信链路、交换设备和信息服务器等
- 与所在领域的垄断企业的关系

5.3.1　服务

当代电信运营商运营提供多种类型的服务，例如，电话服务和因特网服务。**服务**（service）可以根据不同的层和组分类。图5-2只列出了其些层和组，或者是主要的层和组。但是，即使这一图表并不太完整，它也显示了当代电信服务的范围，以及它们之间关系的复杂性。根据提供服务的网络的种类（电话网络，计算机网络），一些服务被结合成组。如果你想画一张更完整的图表，你可以加入电视网络和广播网络所提供的服务。

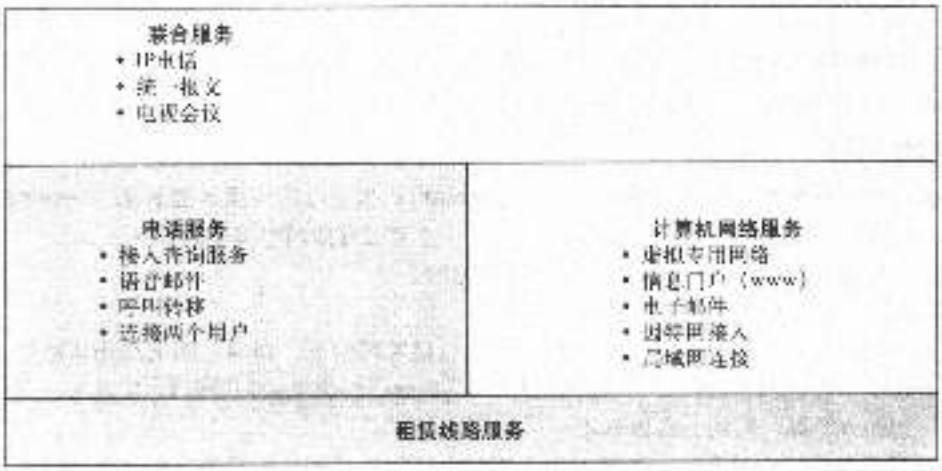


图5-2　电信网络提供的服务类型（深色部分对应于电信运营商提供的传统服务）

租赁线路服务是最低层的服务，因为使用这一服务的用户需要基于这些租赁的线路，建造它们自己的网络基础结构。在它们从这项服务获益之前，它们必须安装电话交换机或分组交换网络交换机。通常，使用这项服务的客户或者是没有自己的通信链路的电信运营商，或者是基于这些所租赁的线路，建造私有的、企业范围网络的大公司。企业范围的网络将在下一小节介绍。

下一个层由两大服务组构成——**电话服务**（telephony service）和**计算机网络服务**（computer network service）。电话服务和租赁线路服务是传统的服务，很久之前就已出现。

计算机网络服务比电话服务出现晚得多，即使今天，就收入而言，它们也远远落后于传统的电话服务。但是，由于它们前景极佳，发展速度也远远领先于传统的服务，大多数电信运营商提供了计算机网络服务。就绝对数量而言，数据流量已经超过了语音流量。但是，数据传输服务的低速率依然使它们无法在收入上跟上传统的服务。

之前介绍的每一层服务都可以被分为若干子层。例如，基于因特网接入服务（将计算机或局域网连接到因特网上），运营商可以提供客户组织VPN的能力，这将使客户比其他因特网用户更安全，或者，运营商可以提供给客户创建门户网站的能力，并将这一网站放在服务提供商的网络中。

目前，服务层次的最高层由联合服务占据，它们基于计算机网络和电话网络的协调运作而实现。这一服务的重要例子是国际IP电话服务，它从传统国际电话服务中抢走了许多客户。

联合服务是网络融合的直接结果，也是这一过程的驱动力。

服务也可以根据**运输服务**（transport service）或**信息服务**（information service）而分类。根据这一分类，电话通话是运输服务，因为电信运营商在用户和用户之间传输语音流量。电话网络的查询服务是信息服务的一个例子。

这种电信运营商提供的服务类型的差异也反映在它们的名字上。主要商业领域为电话服务和租赁线路服务（即运输服务）的传统公司通常被称为**运营商**（carrier）。术语**服务提供商**（service provider，SP）随着因特网和WWW服务（一种信息服务）的爆炸式发展而变得流行。

服务不但可以根据它们提供的信息类型加以分类，还可以根据它们交互性的程度来分类。因此，电话网络提供**交互服务**（interactive service），因为两名用户参与通话（如果是电话会议的话，就有多位用户），并且不时地进行交互。计算机网络也提供了类似的服务，用户可以一边浏览网站的内容，一边回答注册表格的问题，或者，玩互动游戏。

在另一方面，广播网络和电视网络提供非交互的**广播服务**（broadcast service）：信息只在一个方向上传输——根据从一点到多点的安排，从网络到用户。

5.3.2 客户

信息通信服务的消费者可以被分为两大类：**大众客户**（mass individual client）和**企业客户**（corporate client）。

在第一类中，对客户的服务驻地通常是他们的住宅，客户是那些需要诸如电话通信、电视、广播、因特网接入等基本服务的普通居民。对于这样的客户，最重要的因素是服务的经济性：较低的月租费；可以使用标准的终端设备，例如电话、电视机、PC；可以使用诸如双绞线或电视同轴电缆这样的已有线路。复杂的、难以使用的、昂贵的终端设备，例如计算机化的电视机或是IP电话，不太可能被广泛接受，除非他们的价格和传统的电视机或电话相差不大。

除了价格因素，这样的设备还必须支持简单的用户接口，不需要用户参加特殊的课程才能掌握这些设备的使用。大多数建筑中的现有线路，也成为了因特网接入和计算机网络提供新服务的一个限制，因为这些线路起初并不是为数据传输而建设的。安装新的高质量线路（例如光

纤）非常昂贵。正因为这样，家庭用户更多地使用低速的拨号调制解调器连接到因特网上。但是，现在，诸如数字用户专线（DSL）这样的新技术变得更加流行，使得通过已有电话线路的数据传输比传统调制解调器要快得多。除此之外，使用有线电视网络进行数据传输的接入技术也已经存在。

企业客户（corporate client）通常是各种组织机构或公司。如果你从所需要的服务来看的话，小公司和个人用户并没有显著的区别。通常，这些服务是相同的基本电话服务、电视服务、标准的拨号接入因特网信息资源的服务。唯一的差别是所需要的电话号码的数量（通常，两个或更多）。

大型公司由分布在不同地理位置的部门或附属公司构成，并且有经常在家里工作的移动用户（telecommuter），这样的公司需要一整套扩展的服务。这些服务必须包括VPN。提供这一服务的运营商提供一种途径，使得公司觉得它的所有部门和下属机构都由私有网络连接起来（即，网络完全由客户自行拥有和管理）。因此，必须用到运营商的网络。运营商的网络是公共的网络，同时传输许多客户的数据。

现在，企业客户不但需要运输服务，还常常需要信息服务。例如，他们将他们的网站和数据库移到服务提供商处，后者承担一定的职责，包括维护它们的运作、保障企业的雇员可以快速访问这些信息资源，有时候，也需要允许运营商网络的其他用户访问这些资源。

5.3.3　基础结构

基础结构（infrastructure）：除了客观原因影响了电信运营商提供的服务之外，技术因素也扮演了一个重要的角色。为了提供线路租赁服务，运营商必须拥有它自己的传输网络。举一个例子，让我们来看一下SDH，或是诸如ISDN的电路交换网络。为了提供信息服务，需要创建连接在因特网上的网站，这样因特网用户就可以访问它们。

如果运营商并不拥有提供特定服务的整个基础结构，可以使用另一个运营商的服务。把这样的服务和运营商自己的基础结构中的一些元素结合起来，便可以创建出客户需要的服务。例如，一个电信运营商被要求创建一个公共的电子商务网站，但它可能没有连接到因特网的IP网络。为了提供这一服务，运营商可以创建出内容，然后将它放置在另一个运营商的计算机上，后者的网络连接在因特网上。出于创建电话网络或计算机网络的需要而进行的物理通信链路租借，这是硬件或软件基础结构某个元素缺失，而要提供服务的另一个例子。为另一个电信运营商提供服务的运营商通常称为“运营商的运营商”。

在大多数国家，电信运营商必须获取牌照，以提供特定类型的服务。但情况并不总是这样。事实上在所有的国家，就国家的范围而言，运营商都是电信服务市场的垄断者。现在，电信服务反垄断的过程正在快速进行。

美国电信市场的反垄断

通常，垄断者会失去他们的特权。有时候他们会被强行分成小公司。例如，在美国，直到1984年，AT&T还是本地电话服务和长途电信的垄断者。在1984年，根据法庭的决定，AT&T被分为几个小部分，其中最重要的是AT&T长话部和23个贝尔运营公司（BOC），前者只允许提供长途服务，后者只能提供本地电话服务。为了提供地区性的服务，ROC联合起来形成了七个地区性的BOC（RBOC）。

全国性的垄断者被剥夺了他们的特权，并分割为多个小公司，现在他们必须和在本地服务市场、地区性服务市场和长途服务市场出现的新运营商争夺客户。这些运营商通常称为竞争性运营商。在美国，电信服务市场的竞争性发展在1996年加速，这一年，国会批准了电信法案，废除了电信运营商只能在一个细分市场（或是长途/

区域性，或是本地服务）提供服务的限制。现在，美国竞争性本地交换运营商（competitive local exchange carrier，CLEC）非常多，他们就像之前的传统本地交换运营商（incumbent local exchange carrier, ILEC）。在美国地区性和长途市场上，竞争也很激烈，那些市场里有许多称为交换运营商（Interexchange Carrier，IXC）的大型运营商。由于这些术语有时候用来描述项目解决方案和技术，所以它们比较重要。也正是因为这样，在由运营商、服务提供商以及提供的特定服务组成的系统中，电信运营商的类型（IXC、CLEC或ILEC）表示了运营商的位置。

5.3.4 覆盖的范围

根据运营商提供的服务的覆盖范围，运营商可以被分为本地运营商、区域运营商、国家运营商和国际运营商。

本地运营商（local carrier）在一个城市区域或乡村区域运营。**传统本地运营商**（traditional local carrier）（在美国的术语中，称为ILEC）是城市电话网络的运营者，这样的电话网络拥有所有的运输基础结构：将用户驻地（公寓、建筑、办公室）连接到中央局的本地回路。在中央局之间有电话交换机和通信链路。目前，除了传统本地运营商外，还有一些**竞争性运营商**（CLEC），它们常常提供新类型的服务，主要和因特网相关。有时候，它们也在电话领域和本地运营商竞争。

尽管在电信领域开展了反垄断，但传统的本地运营商依然是本地回路的拥有者。

在这种不平等的条件下，竞争性本地运营商常常遇到商业上的困难。它们有几种选择。首先，它们可以只专业化地提供与数据传输和处理相关的附加服务：因特网访问、托管客户的信息资源等。为了组织用户对这些资源的访问，这些运营商需要和传统运营商签订合同，将传统运营商网络中的用户数据转发到竞争性运营商的网络中。这时，服务提供商专业性自然而然地就清晰地显示了出来，因为每一个服务提供商总是在它们公司基础结构最适合的领域专业化。在这样的条件下，合作意味着新的服务。第二，竞争性运营商可以从传统运营商处租借本地回路。通常，传统运营商不太愿意这么做，虽然某些国家的立法机构鼓励甚至强制要求这类行动。第三种选择是建立自己的本地回路网络。这时，竞争性运营商有两种选择：有线本地回路和无线本地回路。基于对私人住宅数量的考虑，由于存在铺设电缆、从本地政府购买牌照等的困难，有线本地回路通常在经济上行不通。这一情形造成了大家对无线解决方案的强烈兴趣，无线解决方案正在快速发展。

区域运营商（regional carrier）和**国家运营商**（national carrier）在很大的范围内提供服务，它们有它们自己的运输基础结构。这一类传统运营商在本地运营商的用户交换机之间传送语音流量。通常，它们拥有连接到高速通信链路的大型中间用户交换机。这些公司通常是运营商的运营商。它们的客户包括本地运营商和大型公司，这些大型公司的部门和附属公司分布在某个区域的多个城市，甚至是多个国家。这类运营商先进的运输基础结构使得它们可以提供长途服务、传输大量的信息。

国际运营商（international carrier）的服务跨越多个国家。其中最著名的是Cable & Wireless、Global One，以及Infonet。它们拥有覆盖多个大洲的主干网。通常，这些运营商和全国运营商合作密切，使用全国运营商的接入网向客户递送信息。

5.3.5 不同类型运营商间的关系

图5-3显示了不同类型运营商的关系，以及它们网络的关系。这张图显示了两种类型的客户——个人客户和企业客户。请记住每个客户通常都需要两种类型的服务——电话服务和数据服务。通常，个人客户的家中拥有电话和计算机。企业客户拥有它们的网络——由PBX支持的电话网络，以及基于公司的交换机而建的、用作数据传输的局域网。

为了连接客户的设备，运营商组织了**接入服务提供点**（points of presence，POP）——它们是放置接入设备的建筑或场地，这些接入设备用来从客户处连接大量本地回路。有时候，这些POP被称为中央局，电话运营商常使用这一术语。用户连接到本地运营商的POP上。本地运营商或大型企业客户需要高速接入和覆盖，以将它们在不同城市和国家的分支结构或办公室连接起来，它们连接到高层运营商的POP上。

由于网络融合的过程还没有创造出能服务各类流量的统一的网络，这张图中每个运营商的网络代表着两个网络——电话网络和数据网络。

从图5-3可以看到，在今天竞争性的电信领域中，并不存在严格的运营商的层次。运营商之间的关系，以及它们网络之间的关系可以很复杂。例如，CLEC2的网络不但与区域运营商3之间有直接的连接，正如层级结构所要求的那样，而且它还和全国运营商3之间有直接的连接。此外，如图5-3所示，不是所有的运营商都拥有运输基础结构（例如，CLEC1）。CLEC1常常只提供额外的信息服务（例如，它向ILEC1用户提供视频点播，或者开发并维护他们的主页）。如图5-3所示，这些运营商也将它们的设备（例如，视频服务器）放置在其他运营商的POP处。

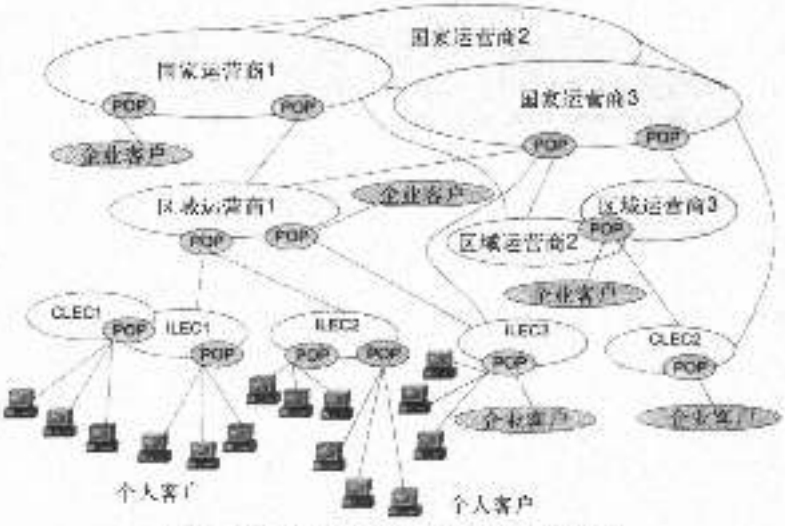


图5-3　不同类型电信运营商之间的交互关系

5.4　公司网络

企业网（corporate network）是一种以支持特定企业运营为目的的网络，企业范围内网络的用户是这个企业的雇员。与电信运营商网络不同，企业范围的网络并不向第三方组织或外部用户提供服务。

虽然任何企业所拥有的、任意规模的网络都可以被认为是企业范围的网络，但是，这一术语通常用来指大型企业拥有的网络，这些企业在不同的城市和国家有部门或附属机构。因此，企业范围的网络通常是由局域网和广域网构成的互联网络。

企业范围网络的结构通常与之前介绍的电信网络结构相一致。但是，它们之间还是有一些区别。例如，连接终端用户的局域网这时候被整合进企业范围的网络中。更进一步，企业范围的网络中，各个结构分支的名称不但反映了所覆盖的地理范围，还反映了公司的组织结构。因此，企业范围的网络常常被划分为部门或工作组的子网、楼宇或校园网，以及主干。

5.4.1 部门网

部门网（department network）是由公司相同部门工作的员工所使用的网络。这些员工在诸如会计部或市场部里，解决共同的问题，或者完成共同的任务。通常假设单个部门可能有100~150名员工。部门网由覆盖部门所有区域的局域网构成。根据不同的情形，这可能是几间房间，或是楼宇的一整层。

部门网的主要目的是共享本地资源，例如，应用程序、数据、激光打印机、调制解调器。通常，部门网有一到两台文件服务器，一些集线器和交换机，以及小于30名的用户（图5-4）。大多数的公司流量位于这些网络内部。部门网通常基于单一网络技术——以太网（或是以太网家族中的几种网络技术——以太网、快速以太网、用得较少的千兆以太网）、令牌环，或是光纤分布式数字接口。通常，这类网络使用不多于两种类型的操作系统。



图5-4 部门网的例子

在部门层和网络维护相关的任务相对比较简单：增加新用户简单故障的恢复、安装新的节点和新版本的软件产品。对这类网络的管理和维护任务可以分配给一名兼职网络管理任务的员工。通常，部门层网络的管理员并不是一位经过特殊训练的专家。然而，这位管理员往往要比其他人更了解一些计算机的软硬件，他/她的任务是完成管理性的工作。

另一种类型的网络类似于部门层的网络——**工作组网**（workgroup network）。这些是具有10~20台计算机的小型网络。工作组网事实上和之前介绍的部门网没有区别。工作组网有一些显著的性质，例如网络的简单性和同质性。

工作组网常常使用基于共享介质的局域网技术。通常，网络的层级越高，共享介质就用得越少，而往往会使用交换局域网。

部门网可以位于楼宇网或校园网内，它们也可以是一个远端办公室的网络。部门网使用局域网技术连接到楼宇网或校园网上。现在，这可能也是以太网家族的代表之一。远端办公室网络通过使用一种广域网技术（例如，帧中继），直接连接到主干网上。

5.4.2 楼宇或校园网

楼宇网（building network）或**校园网**（campus network）将同一公司多个部门拥有的网络连在一起。这些网络位于单个楼宇中，或是几平方英里的范围内（图5-5）。建造这样的网络使用了局域网技术，因为局域网技术的能力足够覆盖这一范围。

通常，楼宇网和校园网根据层级原理建造（即它有它自己的主干，通常基于千兆以太网技术）。连接在主干上的工作组网和部门网通常基于快速以太网或以太网。千兆以太网主干事实上总是交换式的，虽然这种技术的一种变形是基于共享介质的。

这种网络提供的服务包括部门网的互连、访问公司的数据库、传真服务器、快速调制解调器、打印机。因此，每个部门的员工都可以访问其他部门网络上的部分文件和资源。校园网提供的重要服务是独立地接入企业范围的数据库。

只有在园区的层次上，整合不同类的硬件和软件的问题才会出现。每个部门的计算机、网络操作系统、网络设备的类型都可能不同。因此，校园网很难管理和控制。通常，这种规模的网络是基于IP技术的互联网络。



图5-5　校园网的例子

5.4.3　企业范围的网络

企业范围的网络（enterprise-wide network）与众不同之处在于信息服务具有显著的地位。这些网络不能仅仅被限制在运输服务上。电信运营商网络可能提供、也可能不提供信息服务（因为终端用户的计算机在它们的职责范围之外），而企业范围的网络与它不同，它们需要提供信息服务，终端用户的台式机以及服务器都是企业范围网络整体的一部分。参与维护这类网络的开发和支持人员必须考虑这一点。可以说，企业范围的网络是信息电信网络的一个例子，在它之中，两种类型的服务互相之间是平等的。

企业范围的网络的另一个特点是它的规模。部门局域或楼宇局的网络很少被叫做企业范围的网络，虽然从正式意义上说，它们确实是。通常，术语企业范围的网络指代由许多部门范围的网络或楼宇范围的网络组合而成的网络，这些部门范围网络和楼宇范围的网络位于不同的城市，用广域网链路连接起来。

从原理上说，企业范围的网络的结构与图5-1所显示的一般网络结构一致。广域网技术用来将公司的局域网连接起来。这些局域网可能是工作组网络、部门网络、楼宇网或校园网。这些网络由**主干**（backbone）网和**接入**（access）网构成。企业使用和电信运营商同样的广域网技术：ATM或帧中继。IP技术常常被用来连接局域网和广域网，形成企业范围的网络。

在企业范围的网络中，用户和计算机的数量可以数以千计，服务器可能会有数百台。覆盖不同区域的网络之间的距离可能非常远，以致于必须使用广域网链路（图5-6）。为了将远端的局域网或计算机连接到企业范围的网络上，可以使用多种不同的通信工具，包括传输网的电路、无线电信道和卫星通信。

高度的异构性是这类复杂的大规模网络必备的属性。我们无法使用同一类的软件和硬件来满足大量用户的需求。企业范围的网络中有不同类型的计算机，从大型主机到个人电脑；有不同类

型的操作系统，以及大量不同的应用程序。企业范围网络中异构的部分必须像一个整体一样运行，向用户提供他们所需要的所有资源的便捷访问方式。

企业范围网络的出现很好地解决了“从量变到质量”这一著名哲学命题。在将一个大型的、地理位置分散的公司的各个独立网络连成单一网络时，大多数互联网络的性质特点将超出它的关键阈值，之后，它又获得了新的性质特点。在这些条件下，现有的那些解决小型网络常见问题的方法变得不再适用。那些在工作组网络、部门网络、甚至是校园网络中被不足道的问题（或者是还没有出现的问题）变得非常重要。这可以通过一个简单的任务来阐述，例如，支持用户账号信息。

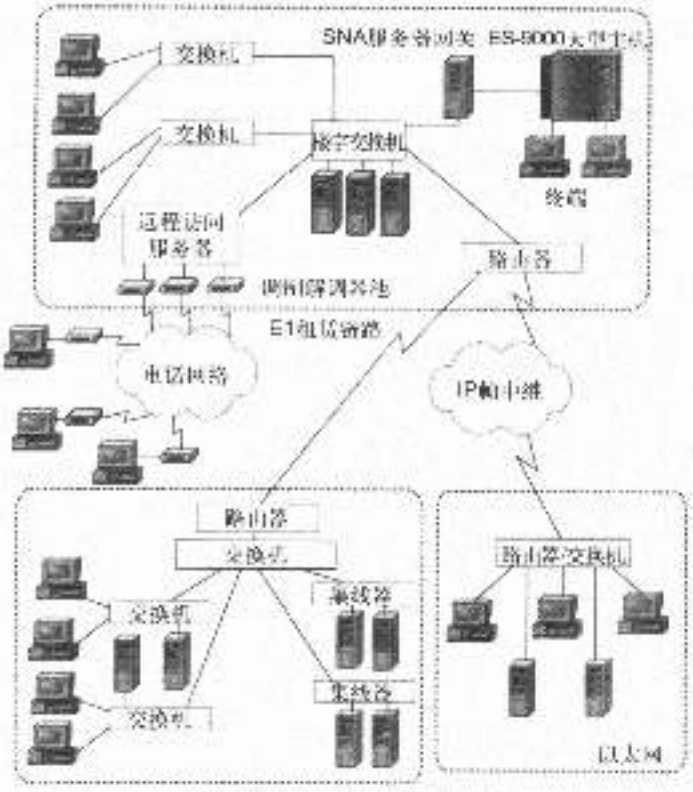


图5-6 企业范围网络的例子

解决这一问题的最简单方法是，将所有用户的账号数据存储在每台计算机的本地用户账号数据库里。当任何用户试图访问这些资源时，所需要的数据从本地用户账号数据库中获取。基于这些信息，访问会被允许或拒绝。对于只有5-10台计算机的小型网络来说，这一方法能很好地工作，但是，如果网络拥有几千位用户，其中的每位用户都需要访问几十台服务器，那么，这一解决方法将变得非常低效。网络中有新服务器时，管理员必须重复进行许多次创建用户账号的操作。此外，当需要访问新服务器的资源时，终端用户必须重复登录的过程。在大规模网络中，这一问题更好的解决方法是使用集中式的目录服务，它的数据库可以储存所有网络用户的账号。管理员只需要完成一次创建用户账号的操作，用户也只需要登录一次网络（请注意，这时候，用户登录了整个网络，而不是单个服务器）。

在从简单网络向复杂的网络，从部门范围的网络向企业范围的网络转移时，地理上的距离变得更大，支持计算机之间的通信也变得更加困难和昂贵。随着网络规模的扩大，对它的可管性、性能、功能的要求也相应地增长。网络中循环的数据量增大，并且，除了这些数据的可获得性，

网络还必须保障它们的安全性。所有这些因素使得企业范围网络的建造基于更强大、更多样化的设备和软件。

5.5 因特网

因特网并不仅仅是一个单一的网络，它是当代文明的一种表现。因特网的出现所带来的变化是多方面的。WWW超文本服务通过将文本、图片和声音等结合在网页中，彻底改变了信息表示的方式。事实上对所有公司（以及通过电话网络的个人）而言，因特网运输都很便宜而且易获得，它极大地简化了建造企业范围网络的任务。与此同时，它也带来一个重要的问题——在支持上百万用户的公共网络上，如何在传输过程中保护企业数据的安全。整个因特网基于TCP/IP栈，它成了最流行的协议栈。

因特网渐渐发展成为用作公共通信的世界范围的网络。它的用途正在不断增加，不但可以用作发布信息（包括推销商品），还可以用来实现商业交易，例如，购买货物和服务，移动金融财产。对于许多公司而言，这意味着它们商业模式的重大变化。此外，因特网也改变了客户的行为，越来越多的人开始喜欢电子交易。

5.5.1 因特网的独特性

因特网的独特性体现在许多方面：

- 就用户数量、覆盖范围、总的传输流量，以及连接的用户数量而言，因特网是世界上最大的网络。虽然在20世纪90年代中期因特网革命后，因特网的发展速度略微降低了，但是，它依然发展迅猛，大大地超过了电话网络的发展速度。
- 因特网是没有单一的控制中心的网络。但是，它根据一些规则运作，提供统一的服务给所有用户。因特网是网络的网络，但是，任何连接在它之上的网络都由一个独立的运营商管理，称为**因特网服务提供商**（Internet service provider，ISP）。虽然存在着一些权威机构，但是，他们只是负责技术政策的统一、协调良好的技术标准集，以及在这样一个巨大的网络中，集中式地指定一些非常重要的参数。这包括连接在因特网上的计算机和网络的名字和地址。但是，他们并不负责因特网每天的维护，或是确保它在一个可用的状态下工作。
- 高度的非集中性有优点和缺点。一个优点是扩展的便捷。例如，为了开始商业运作，新的ISP只要和至少一家现有的ISP签订合同，之后，新ISP的所有用户就可以访问所有的因特网资源。非集中性的缺点包括与因特网技术和服务现代化相关的复杂性。重大的变化需要所有ISP协调一致的努力（需要注意的是，如果网络有单一的所有者，那么这样的改动就会容易得多）。正是由于这些复杂性，许多新的、有前途的技术只是在单一提供商的网络中运用。其中的一个例子就是组播，对于因特网上语音和视频广播的有效组织，这一技术非常必要。至今，这一技术仍然不能克服不同ISP之间的边界。另一个例子是因特网服务相对较低的可靠性。这是因为没有ISP对最后的结果负责，比如说，客户A访问站点B，而A和B属于不同ISP的网络。
- 因特网是一个不昂贵的网络。例如，因特网电话这一新因特网服务之所以能够流行，在很大程度上是因为比起传统的电话网络，它的国际电话通信费用要低得多。更重要的是，这一低价格并不是因为公司为了扩展市场份额而使用的阶段性营销方法。恰恰相反，这一低价格由客观的原因造成，例如，比起传统电话网络的基础结构，因特网运输基础结构具有分组交换网络的特性，它的成本低得多。当然，有人担心随着技术和服务变得越来越先进，因特网会变得越来越贵。因特网技术的开发者和ISP知道这一危险，因此，他们从这一角度检查每一项创新。

但是，如果没有另一个独特的特征，因特网永远不会变成现在的样子——它的大量的信息内容(content)，以及对所有因特网用户而言，对这些内容的易访问性。因特网以网页的形式，存储了数以千兆字节计的，可为终端用户获取的信息。直到1991年，因特网还是一个针对较少的，但是世界范围的用户的流行网络。他们大部分是大学或科研机构的雇员和学生。其他所有需要数据网络服务的客户，例如大公司，银行，政府组织，都使用其他的分组网络，即，X.25。

X.25网络以及在20世纪90年代中期代替它们的帧中继网络，都完全不同于WWW服务。随着Web的出现，用户立即明白，方便而有用的工具出现了。在Web发明之前，因特网主要用作信息系统，而不是运输。因特网刚出现的头几年，最主要的应用是电子邮件和FTP档案。但是，访问FTP档案中存储的文本信息的工具相当简单，因此，根据文件名查询所需要的信息要花几小时甚至几天。

方便地表示信息内容不同部分之间的交互关系（例如超链接）和标准图形浏览器激发了因特网的革命，这些浏览器能方便有效地运行在各种操作系统上。因特网迅速充满了网页形式的信息，渐渐地变为百科全书、日报、促销机构、大商场。现在，许多人无法想像如果不能上网，他们的生活将会怎样——为了和朋友通信、为了查找急需的信息、为了寻找新工作，或是为了支付账单。

注意 但是，如果认为因特网使其他的网络技术不再有用，那就错了。事实并不是这样，而且也不太可能会发生。TCP/IP将技术互连起来，为其他技术提供了空间，即每一个单独网络中使用的技术构成了因特网。因此，因特网的成功并不能构成只学习TCP/IP技术的原则，在当代网络中，TCP/IP和许多其他技术紧密交互，例如，以太网，ATM，帧中继，MPLS和ADSL。

5.5.2 因特网的结构

网站所包含的信息吸引了越来越多的人，这也极大地改变了企业用户和电信运营商对这一网络的看法。现在，几乎所有的传统电信运营商都支持因特网。此外，许多新公司将业务建立在提供因特网服务的基础上。因此，图5-7所显示的因特网的一般结构，在很大程度上已经成为全世界电信网络的一般结构，图5-3已经讨论了其中的一部分。

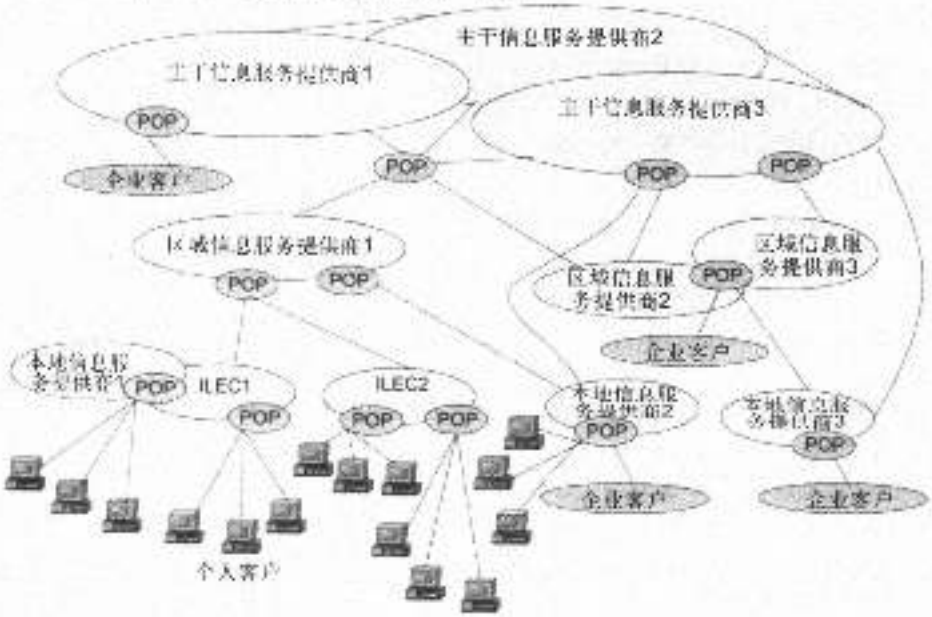


图5-7 因特网的结构

对ISP的分类有许多种方法。图5-7显示了其中的一种，这种方法和对电信运营商的分类方法类似。根据这一分类方法，ISP的主要特性是覆盖范围、地域以及所提供的服务。**主干信息服务提供商**（backbone ISP）类似于国际电信运营商。它们拥有覆盖很大疆域（某些国家、大洲、甚至全世界）的主干网。主干信息服务商的例子包括Cable & Wireless、MCI、Global One这样的公司。相应地，**区域信息服务提供商**（regional ISP）提供了在一定区域范围内（州、国家、地区——取决于特定国家的行政区域划分）的因特网服务，**本地信息服务提供商**（local ISP）通常在一个城市的范围内提供服务。

ISP之间的关系基于为互相之间的流量传输而订立的对等商业协议（*peer-to-peer commercial agreement*）。主干提供商通常与所有其他主干提供商签有这样的协议（由于主干提供商数量有限），区域提供商通常和一家主干提供商及多家区域提供商签有这样的协议。同时，提供商配置它们的设备，以保证流量在两个网络间的双向传输。

网络接入点/因特网交换中心

为了简化对提供商间通信过程的组织，因特网上有特殊的交换中心，那里连接着许多提供商的网络。这些交换中心可以由特定的高层提供商（国家或国际提供商）为低层提供商提供支持。交换中心也可以由一些专门完成这一任务的公司支持。这样的交换中心有特殊的名称——通常是因特网交换中心（IX）或网络接入点（NAP）。

在ISP网络间的流量交换中，NAP/IX可以扮演多种角色。从最小的形式而言，这样的中心简单地提供给ISP一个安装通信设备的驻地。ISP建立起所有它们自己的物理连接和逻辑连接。更为常见的，NAP/IX的通信设备参与ISP之间的流量交换。在这种情况下，NAP/IX只在所有ISP的设备间提供物理连接，而不在ISP的网络间创建逻辑连接提供流量交换。因此，使用这种方式连接到NAP/IX的ISP仍然需要互相之间达成协议。最后，有些数据交换中心将流量交换功能和商业功能结合起来。这些中心称为交易所，它们担任了批发带宽交易的交易所功能。所有连接在这些中心的ISP都会宣布它们数据传输的价格，在达成协定的过程中，中心担任着协调者的角色。

另一种流行的ISP分类方法将它们分为四类——**第一层**（Tier 1）、**第二层**（Tier 2）、**第三层**（Tier 3）、**第四层**（Tier 4）（参见http://www.nwfusion.com）。

第一层ISP、第三层ISP、第四层ISP的定义与之前介绍的主干因特网服务提供商、区域因特网服务提供商、本地因特网服务提供商一致。然而，第二层因特网服务提供商被定义为特殊类型的ISP。

第二层因特网服务提供商向一个国家、甚至整个洲的大量终端用户提供因特网服务。它提供了一系列信息服务和通信服务。第二层因特网服务提供商类似于本地因特网服务提供商，它直接和因特网用户打交道。但是，覆盖范围的规模将它和本地提供商区别开来。诸如American Online这样的公司是第二层因特网服务提供商。

第二层因特网服务提供商的出现，是基于与多个无法独立提供因特网服务的本地电信运营商之间达成的协议。在图5-7中，ILEC2公司是一个这种类型的公司。ILEC2拥有最初用于电话流量的本地回路。现在，它的用户可以使用同样的物理信道通过调制解调器来传输数据。当前有两种类型的调制解调器适用于本地回路——拨号和异步数字用户线（ADSL）。拨号调制解调器暂时地将终端用户的计算机连接到ISP的网络上，类似于在一段通话时间内，电话将用户连接到电话网络上。ADSL调制解调器保证了在计算机和ISP网络之间的持续连接。

由于ILEC2只提供电话服务，它在它的POP中将电话流量和数据流量分开。然后，它通过使用电话交换机，按照通常的方式处理电话流量。同时，它将数据流量转发到与它达成协议的ISP那里（在图5-7中，区域因特网服务提供商1）。如果ISP与许多本地运营商达成了协议，它就成为了没有

自己的供用户接入的基础结构的第一层因特网服务提供商。

通常，第一层因特网服务提供商通过第一层因特网服务提供商与其他的ISP交互。第一层因特网服务提供商在很长的距离上传输流量，并提供其他有用的服务，例如，解决互相的付费问题。

对术语第一层到第四层有多种不同的解释。例如，在某些书里，会发现这些术语的定义只考虑了服务覆盖的地理范围（即它们与国际因特网服务提供商、主干因特网服务提供商、区域因特网服务提供商、本地因特网服务提供商一致）。

另一种对ISP分类的方法是根据它们所提供的服务的类型。在这种情况下，通用的术语ISP通常指那些只向终端用户提供运输服务的公司。也就是说，它们保证了将它们的流量传输到其他的ISP网络。

- 如果一家ISP有它自己的网站，并且提供了内容，那它就被称为**因特网内容提供商**（Internet content provider，ICP）。大多数的ISP也是ICP，因为它们支持了它们自己的信息站点。
- 如果一家公司为另一家公司创建的内容提供了驻地、链路、服务器，那么它就被称为**托管提供商**（hosting provider）。
- 另外还有**内容分发提供商**（content distribution provider，CDP），它们不创建内容，但是在接近用户的多个位置托管内容，以增加用户访问信息的速度。
- **应用服务提供商**（application service provider，ASP）提供用户访问它们所支持的大规模、通用的软件产品。通常，这些用户是对公司管理应用软件（例如SAP R3）感兴趣的企业用户。
- 由于因特网成为了一种社会现象，提供服务的供应商的数量正在不断增加。例如，有些公司与当地政府和取暖电力供应商合作，提供通用账单支付的服务（**计费服务提供商**，billing service provider）。

5.5.3 因特网的边界

现在，让我们来看一看因特网的边界。在读完这节后，你可能会问一个问题：如果因特网被整合到电信运营商的公共基础结构中，是否还可能找出因特网的边界？

为了回答这一问题，让我们来考虑一个典型的ISP网络以及它的客户（图5-8）。

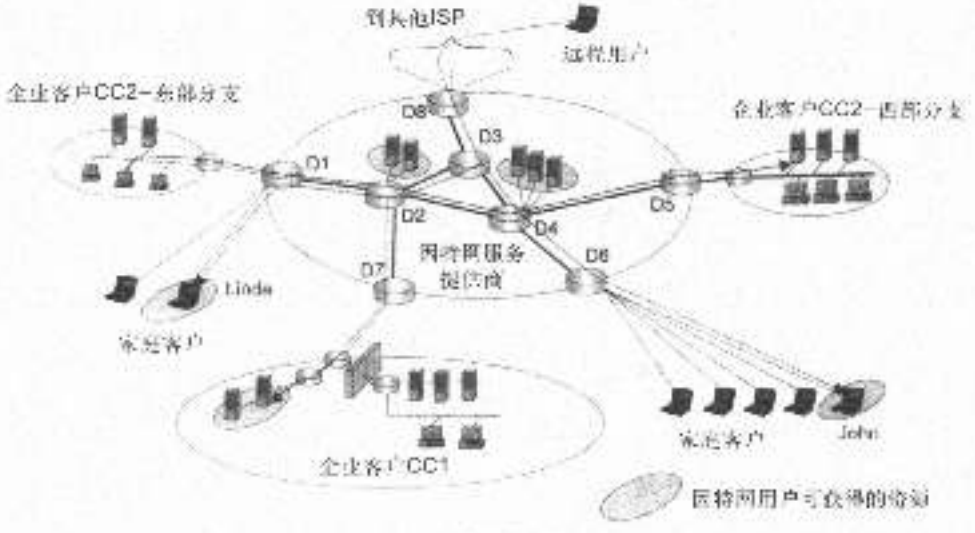


图5.8 因特网的边界

这一例子中的ISP有两个企业客户和大量的个人客户。很自然地，这家ISP还有多条和其他ISP的连接，通过这些连接，它可以访问构成因特网的所有其他ISP。在ISP网络中，多个服务器托管了网站，这些网站可以被所有因特网用户访问。然而，客户CC1有一些保密的公司信息资源。

只有公司的雇员才能访问这些资源。因此，企业客户CC1在它的网络中安装了特殊的通信设备，称为防火墙。防火墙保护了一部分网络，这部分网络含有只供内部使用的服务器。防火墙只对网络内部的用户提供对这些服务器的访问，它阻塞了所有外部客户对这些服务器的请求。防火墙允许内部客户访问外部的信息资源（即因特网资源）。它也将外部服务器的回复传回给企业雇员。如果企业用户不需要访问因特网资源，更简单的做法是从物理上断开内部网络和ISP网络的连接，而不安装防火墙。

企业客户CC1有它自己的网络，这一网络包含了多台服务器和许多公司雇员的计算机。企业客户CC1有一条到ISP的持续高速连接，使用了基本因特网接入服务，保障了企业范围的网络和ISP网络之间数据的交换。ISP网络也作为中间网络，提供和其他ISP网络的数据交换。客户CC1支持多个它自己的Web站点，这些站点可供所有因特网用户访问。

在图5-8中，大多数ISP的个人客户使用调制解调器连接到网络上。两位个人客户——John和Linda——使用ADSL调制解调器与ISP网络有持续连接。John和Linda在他们自己的家用电脑上创建了个人网站，每一个因特网用户都可以连接并使用网站提供的信息。

企业客户CC2有两个分支，坐落于不同的镇——西部分支和东部分支。每一个分支都有自己的局域网。这些网络基于供因特网使用的TCP/IP技术。客户CC2使用ISP将它的分支的网络连接到企业范围的网络上，而不是为了获取因特网接入。因特网接入并不是这一客户所需要的服务。相反，这一公司需要保障高度的数据安全性。正因为这样，它和客户CC1不同，CC2公司使用了ISP提供的VPN服务来保护它自己的资源。VPN服务使客户CC2的两个网络和因特网流量隔绝。因此，这两个网络的所有信息资源——网站、数据库等——只能供内部用户使用。CC2公司的雇员不能使用因特网服务，因为它们的请求不会被传输到因特网上。

现在让我来回答你的问题。ISP网络是否是因特网的一部分？一方面，你可以回答是，因为它将客户流量传输到因特网上，并包含了公共信息资源。图5-8所示的远程因特网用户可以请求访问ISP中的任何网站。

另一方面，部分ISP运输基础结构与因特网并不相关。这些网络设备包括通信设备D5和连接D5和D4的链路，前者用来支持客户CC2的运作。除此之外，ISP网络包含了仅仅是部分服务于因特网客户的设备和链路。它们是设备D1、D2、D4以及连接这些设备的链路。

正因为这样，是的回答并不是准确的情形。ISP网络的边界和因特网的边界并不一致，即使因特网是由ISP网络构成的。

相似的情形也存在于客户CC1的网络中。一方面，可以说它的网络不是因特网的一部分，因为这一公司并不是一家ISP，相反，它是一家客户。此外，CC1网络包含了防火墙保护的资源，这些不是因特网的一部分。另一方面，CC1网络包含了可供所有因特网用户（包括远程用户）访问的Web站点。从远程用户的角度看，这些站点和ISP网络中的站点没有区别。CC1网络也包含了防火墙保护的资源。很自然地，这些防火墙保护的资源不是因特网的一部分。

通过拨号连接连到因特网的家庭客户通常不会将他们的Web站点放在家用电脑上，因为在大多数时间里，这些站点不能被其他因特网用户访问。但是，拥有通过ADSL调制解调器进行持续连接的家庭用户可以这么做。在图5-8的网络中，John和Linda支持了他们自己的Web站点，远程用户可以在任何需要的时候使用这些站点的内容。

从用户的角度看，因特网提供了一系列分布在不同网络上的信息资源——ISP网络、企业范围的网络、家庭网络以及个人用户的计算机。

因特网的运输工具也是虚拟的。可以认为它们是电信运营商的部分资源（通信设备和链路），用以保障因特网流量传输——即，因特网用户和信息资源之间的流量（或是，就电子邮件或因特网电话流量来说，两个因特网用户之间的流量）。

总之，ISP的网络通常称为私有IP网络，因为在使用这种网络时，运营商通常既提供因特网服务，也提供其他类型的服务，例如VPN。如果这通过使用因特网所基于的技术而实现（即TCP/IP运输和WWW信息服务），这样的服务就称为**局域网服务**（intranet service）。

小结

- 计算机网络提供两种类型的服务：信息服务和运输服务。通常，术语网络服务（*network service*）被理解为运输服务，是因为考虑到信息传输是网络的主要功能。信息服务由网络的终端节点（服务器）提供。运输服务由中间节点（网络交换机和路由器）提供。
- 计算机网络可以用适用于任何电信网络的一般结构来描述。这个一般结构由以下部件组成：接入网（*access network*）、主干（*backbone*）、数据中心（*data center*）。
- 创建电信网络来提供公共服务、拥有这些网络、支持它们运作的专业公司称为电信运营商（*telecommunications carrier*）。
- 各电信运营商之间的区别在于它们提供的服务、提供服务所覆盖的地理范围、服务所针对的客户以及运营商拥有的基础结构。这些包括链路、通信设备和信息服务器。专业提供计算机网络服务的电信运营商通常称为服务提供商。
- 企业范围网络的主要目的是支持拥有这一网络的企业的运营。企业范围网络的用户是企业的雇员。
- 因特网是一个在全世界范围内提供各种服务的独特的计算机网络。因特网使用TCP/IP栈来互联基于多种不同技术的网络。随着因特网以及它的信息服务（电子邮件、网站、聊天）的流行，TCP/IP栈的运输协议成了用来建造互联网络的协议。

复习题

1. 哪一个术语对应于以下的定义：“用来汇集来自终端用户的设备，通过多个链路到达的信息流的网络”？
   A. 主干
   B. 接入网
   C. 核心网
   D. 运营网
2. 请举出不同类型电信网络中的数据中心的例子。
3. 请列举出接入网和主干网的主要需求。
4. 请列举出电信运营商可能有的客户类型。
5. 什么时候运营商网络可以被称为企业网？
6. 电信运营商网络的主要特点是什么？
7. 对于电信运营商来说，租赁线路服务是一种传统的服务，还是一种较新的服务？
8. 为了吸引客户，刚刚开始运营的竞争性运营商会提供哪些额外的服务？
9. 第一层因特网服务提供商和第二层因特网服务提供商之间的区别是什么？
10. 哪种类型的服务提供了因特网接入服务？
11. 电信运营商是否可能不拥有任何通信链路，而提供因特网接入服务？
12. 请在对网络的描述和它们的类型之间建立对应关系（有一种网络类型并没有描述）。

网络类型	企业网	校园网	部门网	电信运营商网
由一组用户使用(100~150名用户)。所有的网络用户都解决一个特定的商业任务。基于单一的一种技术。				
数千台计算机，数百台服务器。计算机、通信设备、操作系统、应用程序的高度异构性。使用广域网链路。				
在一栋楼宇的范围内连接更小的网络。没有广域网链路。提供所有员工访问公司数据库的服务。				

13. 哪种类型的网络（企业范围的网络还是ISP网络）更多地共享局域网？
14. 从层次上说，企业范围的网络可以被分为哪些层？
15. 请列出ISP专业化的类型。
16. 如果公司网有到因特网的持续连接，这是否表明它是因特网的一部分？

练习题

1. 对于竞争性本地运营商来说，它如何向个人客户提供对它的网络资源的访问？
2. 电信行业反垄断需要解决哪些问题？
3. 请描述对电话用户的虚拟专用网络（Virtual Private Network）服务。
4. 如果公司想成为ISP，并开始向它的客户提供服务，公司的管理层需要按顺序做哪些事？
5. 一家ISP拥有主干网和接入网，它将一个新的数据中心接在哪种网络上更好？

第6章 网络的特性

6.1 引言

计算机网络是复杂和昂贵的系统，它们实现重要的商业任务，服务许多用户。因此，不但需要保障网络的运行，而且还要确保网络运行的可靠性和高质量。

服务质量（quality of service）的概念可以在一个更广泛的程度上理解。这一概念可以包括网络所有可能的性质，以及终端用户所期望的服务提供商的性质。为了使用户和服务提供商可以讨论服务问题，并在一个形式化的基础上建立关系，需要有一些大家共同接受的网络特性。在这一章中，我们将介绍与网络运输服务质量相关的特性。比起信息服务的质量，它们更容易被标准化。运输特性反映了诸如性能、可靠性和安全等主要的网络性质。

在向用户提供服务的过程中，可以对部分特性进行量化的评价和衡量。用户和服务提供商可以达成所谓的**服务水平约定**（Service Level Agreement），确定对某些特性的量化的需求，例如，服务的可用性。

服务质量的概念常常在狭义的意义下使用，作为一种当代网络技术的方向，这时候我们使用缩略词QoS。这一方向的目标是开发一些方法，来保障使用网络进行高质量的流量传输。QoS的特性有一个共同特点——它们都反映了流量传输中排队机制的负面效果，例如，暂时的流量传送速度下降、分组递送的不同延迟，以及由于交换机缓存负载过多而造成的分组丢失。保障QoS的方法将在下一章中介绍。

6.2 特性的类型

6.2.1 主观质量特性

如果我们找几位用户，让他们谈谈对网络服务质量的理解，我们可能会得到许多不同的答案。最可能遇到以下的观点：

- 网络运行得很快，没有延迟。
- 流量可以可靠地传输。
- 可以提供持续的服务。
- 支持服务和帮助热线工作良好，能提供有用的建议，真正有助于解决问题。
- 可以根据一个灵活的计划提供服务。任何时候，网络访问的速度可以在一个很大的范围内增加。
- 服务提供商不但能传输我的流量，还可以保护我的网络免受病毒的攻击和入侵。
- 让我可以在任何时候检查我的流量的传输速度，是否发生数据丢失。
- 除了标准的因特网接入外，服务提供商还可以提供大量的辅助服务，例如，托管我的个人网站和IP电话服务。

这些主观的评价反映了用户对网络服务质量的希望。客户（用户）是任何商业活动最重要的部分，包括数据网络。但是，网络中还有另外一个部分——服务提供商（如果网络是公共的，指商业性的服务提供商；如果网络是属于公司的，指非商业性的服务提供商）。为了使用户和服务提供商都能客观地评估网络的服务质量，网络服务质量的形式化特性（*formalized characteristic of the*

quality of network services）可以对特定的质量进行量化的衡量。

6.2.2　网络特性和要求

使用网络**特性**（characteristic），用户可以形式化地表达对网络的特定**需求**（requirement）。例如，用户可以表示需要如下需求：*我的信息在网络上传输的平均速度必须不低于2Mb/s*。这时，用户使用了被称为“数据在网络上传输的平均速度”这一特性，并且定义了这一特性的值域，这一值域对应于对用户来说良好的服务质量（即网络的有效运行）。

运输服务所有的QoS特性都可以被归类为以下几组中的一组：

- 性能
- 可靠性
- 安全性
- 仅限提供商

前三组对应于运输服务的性质，这些对用户来说最为重要。网络必须以特定的速率传输信息（性能），服务不能有丢失或延迟（可靠性），并且保护它们免遭未授权的访问或篡改（安全）。

自然，服务提供商为了使用户满意，会关注所有对用户而言重要的特性。同时，还有一些网络的特性对服务提供商来说非常重要，但对用户而言无关紧要。

网络服务于大量的用户，服务提供商必须有效地组织它们的运行，来同时满足所有用户的需求。通常，这是一个很困难的问题，因为主要的网络资源——链路和交换机/路由器——共享给用户信息流。提供商必须在所有同时存在的多个流之间，寻找资源分配的平衡，以满足所有用户的需求。对这一问题的解决方法，包括在传输用户流量时，对资源的使用进行计划和控制。因此，提供商对那些描述资源性质的特性非常感兴趣，这些资源用来服务网络的客户。例如，提供商对交换性能很感兴趣，因为提供商必须衡量交换机可以服务的流的数量。另一方面，终端用户对特定交换机的性能并不感兴趣，他们感兴趣的是最终结果——他们的信息流是否得到了高质量的服务。

第四组包含了那些只有服务提供商感兴趣的QoS特性。其中的一个例子是网络的可延拓性（即不改变网络技术的条件下，增加用户数量的可能性）。

6.2.3　时间尺度

在介绍QoS的特性和保障QoS方法之前，熟悉另外一种分类方法可能会有用——这些特性定义的时间尺度，以及QoS方法工作的时间尺度。

长期特性（Long-term characteristic）用来定义几个月到几年的时间段。这些特性可以被称为项目解决方案特性，保障它们的方法是*网络设计和规划*。这一组特性包括了项目的解决方案，例如，选择交换机的类型和数量，选择网络的拓扑结构和通信链路的带宽。这些参数直接影响了QoS特性。有些项目的解决方案是成功的、良好平衡的，保证了网络不会发生拥塞。另一些解决方案可能不太有效，造成了流量的瓶颈，所以，延迟和分组的丢失超过了上限。

显然，对网络设备的整体替换和大规模升级是耗时耗力的工作，通常会花费相当多的金钱。因此，这些工作不会常常发生，它们会在很长一段时间内影响QoS。

中期特性（Medium-term characteristic）用来定义从几秒到几天的时间段。这类特性的例子包括流传输的平均速率，或是分组递送的平均延迟，这些由相当长的时间间隔来定义，在这段时间内，大量的分组被处理。确定路由的方法是这一范围内方法的一个例子。如果网络的拓扑和流量参数维持不变，并且数据链路和网络交换机不失效的话，路由可以几小时或几天维持不变。

短期特性（Short-term characteristic）由处理单个分组所对应的时间来定义（即以毫秒或微秒为单位）。这一组包含了诸如缓存时间、或是单个分组在交换机或路由器队列中所花费的时间这样的特性。专门用来分析和保障这一组特性的方法被称为**拥塞控制**（congestion control）和**拥塞**

避免（congestion avoidance）。在后面我们会更详细地介绍。

6.2.4 服务水平约定

合同或协议是服务提供商和它们的客户（终端用户）之间协调的基础。公共数据网络的服务提供商和它们的客户总是达成某种形式的协议。但是，这样的协议并不总是标明所提供服务的量化要求。通常，这样的协议过于笼统地标明所提供的服务（例如，因特网接入）。同时，还有另外一种类型的协议，通常称为**服务水平约定**（Service Level Agreement，SLA）。在这类协议中，服务提供商和客户用量化的方式描述了所提供的服务质量，使用了网络效率的常用特性。

例如，SLA可能要求服务提供商以客户发送数据到网络同样的平均速率，无损传输客户的流量。SLA也可能规定，如果客户流量的平均速率不超过一个特定的值（例如，3Mb/s），那么当前的协议依然有效。不然，提供商有权丢弃超出的数据。为了使这一协议更确定，每一方都能控制它的遵守情况，就需要标明衡量平均流量速率的时间段（日、小时或秒）。当用来衡量网络特性的工具和方法都被明确标明后，SLA就变得更加确定，这样，服务提供商和用户都可以清楚地理解协议。

SLA不能只由商业服务提供商和它们的客户决定，还需要在许多大公司之间决定。对于后者，SLA在同一家公司的不同部门（例如，IT部门和电信部门）和网络服务提供商间、用户和公司的功能性部门（例如，生产部门）间决定。

6.3 性能

你已经知道了网络设备的主要特性——链路带宽和通信设备（例如交换机、路由器）性能。这些是只有服务提供商感兴趣的网络资源的长期特性。知道了这些特性，提供商确定了它们最多可以服务多少用户，从而计划它们的商业活动。

但是，用户对其他的性能特性感兴趣，这些性能特性可以让他们定量地衡量流量速率和传输质量。为了定义这些特性，我们从考虑理想的网络如何传输数据开始。

6.3.1 理想的网络

如果我们认为一个网络是*理想的*（*ideal*），那就是说，它以一个恒定的延迟传输每一个比特的信息，这一延迟等于光在物理介质中的传播速度。此外，不存在具有无穷带宽的网络链路：链路带宽是有限的；因此，信息源以某个有限的时间间隔，向网络传输分组，而不是立刻就能将分组传送到网络上。这一有限的时间间隔，正如你已经知道的，等于分组大小（以比特衡量）除以链路带宽的商。

图6-1显示了这一理想网络传输分组的结果。上面的轴显示了源节点向网络传输分组的时间；下面的轴显示了分组到达目的节点的时间。换句话说，上面的轴显示了网络的负载，下面的轴显示了这些流量通过网络传输的结果。分组的出发时间是分组的第一个比特传输进网络的时间，分组的到达时间是分组的第一个比特到达目的节点的时间。

从图上我们可以看到，理想的网络：

- 将所有的分组递送到目的节点，不丢失分组，不损毁分组。
- 与所有分组发送时相同的顺序递送所有的分组。
- 以最小的延迟递送所有的分组（$d_1=d_2$，依此类推）。

相邻分组之间所有间隔保持不变非常重要。例如，在发送时，第一个分组和第二个分组之间的间隔等于τ_1，当这些分组到达目的节点时，这一值保持不变。

以最小的延迟和固定的分组间间隔，可靠地递送所有分组，这将满足任何网络用户的需求，无论在网络上传输的是何种类型的流量——不管是Web服务还是IP电话流量。

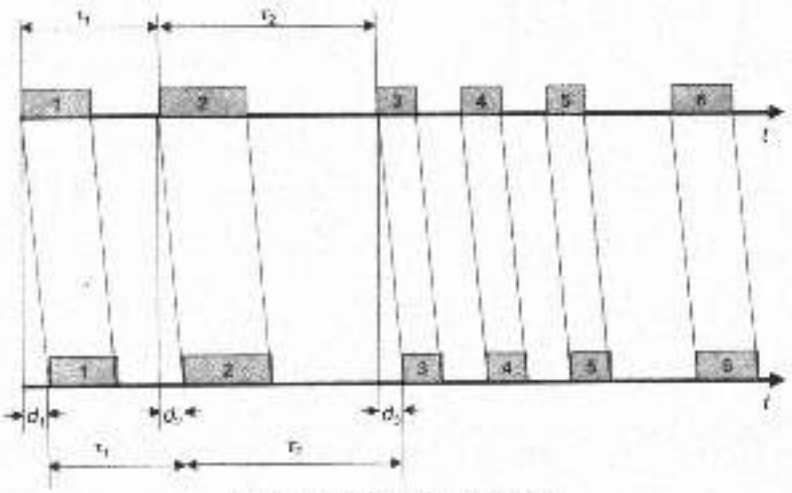


图6-1　理想网络中的分组传输

现在，让我们来看看，在真实的网络中，我们可能会遇上哪些与理想模型不符的地方，哪些特性可以用来描述这些不符（图6-2）。

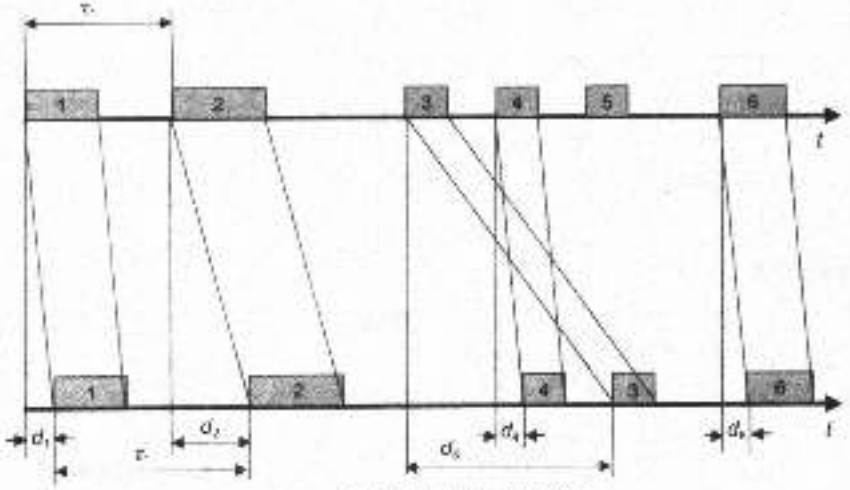


图6-2　真实网络中的分组传输

- 分组以可变的延迟，递送到目的节点。正如你已经知道的那样，这是分组交换网络很常见的特性。排队过程的随机性会造成可变的延迟。同时，某些分组传递延迟可能相当大——比平均延迟值大几十倍（$d_1 \neq d_2 \neq d_3$，依此类推）。因此，相邻分组的时间相关性改变了，这又会反过来造成某些应用的严重问题。例如，在数字语音传输时，分组间间隔的不平均将造成声音的组业。
- 分组可以以一种不同于发送时的顺序，递送到目的节点。例如，图6-2中的图表显示了，分组4在分组3之前到达了目的节点。在数据报网络中会遇到这类问题，同一个流的不同分组通过不同的路由来传输。所以，这样的分组以不同的延迟程度在队列中等待。显然，分组3经过了一个（或多个）拥塞的节点。因此，它的总延迟非常大，以致于分组4比它先到达。
- 分组可能会丢失或损毁。后一种情形等同于分组丢失，因为大多数协议无法恢复损坏的数据。在大多数情况下，协议只能通过帧校验序列（FCS）检测出发生了数据的损毁。

- 目的节点输入处信息流的平均速率可以不同于源节点发送给网络的信息流的平均速率。这更多的是因为分组丢失，而不是分组延迟。因此，在图6-2所示的例子中，由于分组5的丢失，输出流的平均速率降低了。丢失或被毁的分组越多，信息流的速率越低。

显然，每一个分组传输时间的这一系列值，给流量传输质量带来了非常复杂的特性。然而，这些网络性能的特性太宽泛太冗余。

为了以简洁的形式表示QoS特性，我们需要使用统计方法（statistical method）。统计性特性揭示了网络行为的某些趋势。这些趋势只有在很长一段时间内才会变得非常明显。只有追踪数百万分组的传输，网络行为中的趋势才可能被揭示出来。在当代网络中，单个分组的传输时间在微秒的时间范围内。例如，在快速以太网中，这一传输大约花100μs；对于千兆以太网，大约是10μs；对一个ATM信元，从几微秒（ms）到3ms不等（取决于传输速率）。因此，为了得到更稳定的结果，我们需要监测网络几分钟，或者，更好的是，几个小时。这样的时间间隔才能被认为是长的。

有两组与网络性能相关的统计特性：

- 分组延迟的特性
- 信息率的特性

6.3.2 分组延迟的特性

随机变量的分布直方图（distribution histogram）是最主要的统计工具。在这里，需要被衡量的随机变量是分组传递延迟。

假设我们已经测量了每一个分组的传递延迟，并保存了这些数据。为了得到分布直方图，我们需要将所有可能延迟的范围分成几个区间，在每个区间内，计算序列中多少分组属于这一区间。于是，我们得到了图6-3所示的直方图。在这一例子中，所有的延迟值都在25ms到75ms的范围内。网络有一个25ms的固定延迟，这与信号传播和分组缓存有关。我们将这一范围分成六个区间。由此，我们可以使用以下六个数字来表示网络的特性：n1、n2、n3、n4、n5、n6。这一表示形式要精简得多。如果我们将整个值域分成越少的区间，描述测量的值的数量也越少。但是，我们需要在尽可能减少区间数和特征的信息密度之间取得良好的平衡。

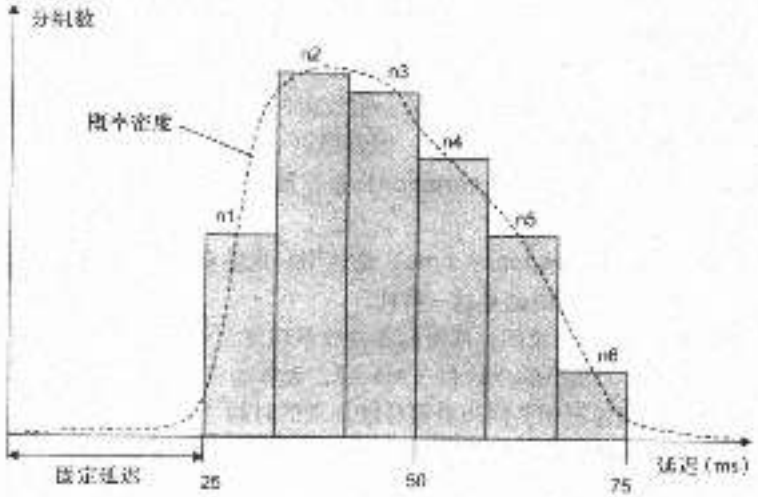


图6-3 表示延迟分布的直方图

延迟直方图很好地表示了网络的性能。通过使用这张图，我们可以衡量哪些延迟是可能发生

的，哪些不太会发生。这意味着我们可以更准确（具有更大的概率）地预测网络的行为。例如，使用图6-3所示的图表，我们可能会说，在0.6的概率下，分组传输延迟不会超过50msec。为了评估这一说法，我们算出传输延迟小于50msec的所有区间内的总分组数，将它除以全部的分组数。换句话说，我们找出了传输延迟不超过50msec的分组的百分数。

如果我们增加区间的数量和监测的时间，到极限的时候，我们将得到一个连续的函数，称为**分组延迟的分布密度**（distribution density of the packet delay）（在图6-3中，它由一条虚线来表示）。从概率论我们可以知道，为了确定随机变量在一个特定范围内取某一个值的概率，我们需要找出这一函数在特定范围下界和上界之间的积分。换句话说，我们需要找出由分布曲线和X轴在一个特定范围内图的面积。

分组交换网络最重要的特点是，这类网络的许多特性体现出统计（概率）特性。我们无法保证在任何给定的实例中，这类特性有任何预定义的值。我们只能说，这些事件有一定的概率，因为分组交换网络中的数据传输过程从本质上说是随机的。

让我们再来考虑一下其他几个常常使用的延迟特性：

- **平均延迟**（average delay，D）表示为所有延迟（d_i）的和，除以所有测量的总数量（N）：

$$D = \frac{\sum_{i=1}^{N} d_i}{N} \tag{6.1}$$

- **抖动**[⊖]（jitter，J）代表了对于平均延迟而言，延迟的平均偏差：

$$J = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - D)^2} \tag{6.2}$$

平均延迟和抖动都用秒来测量。显然，如果所有延迟的值d_i都相等，那么$D = d_i$且$J = 0$（即，不存在抖动）。

- **变动率**（coefficient of variation，C_v）。这是一个没有单位的值，等于抖动与平均延迟之比：

$$C_v = \frac{J}{D} \tag{6.3}$$

变动率刻划了流量的特性，而没有与时间尺度的绝对值相联系。理想的同步流（即，stream）标准差总是0。如果变动率等于1，那么流量是突发性的。

- **最大延迟**（maximum delay）是预定义的概率下，分组延迟不能超出的值。通常，我们使用延迟直方图来决定这些值。为了得到一个可以作为网络运行质量证据的评估，我们有理由指定一个高概率，例如，0.95或0.99。如果一家公司告诉用户，网络保障延迟水平为100msec的概率为0.5，这很可能不能令用户满意，因为用户不知道一半的总分组数的延迟程度。
- **最大延迟变动**（maximum delay variation）是在预定义概率下，延迟偏离平均值所不能超过的最大值。
- **网络响应时间**（network response time）是从用户角度看，网络的整体特性。用户说网络在某一天运行得特别慢，指的就是这一特性。

响应时间是用户向网络服务请求的生成时间和接收到回复之间的时间段。

网络响应时间可以表示为几个部分的和（图6-4）。大体而言，它包含了客户计算机上请求生成的时间（$t_{client1}$）、从客户机通过网络向服务器传输请求的时间（$t_{network1}$）、服务器对请求的处理时间（t_{server}）、从服务器通过网络向客户机传输回复的时间（$t_{network2}$）、在客户机处理服务器回复的时间（$t_{client2}$）。网络响应时间在整体上刻划了网络。它取决于服务器硬件和软件运行的质量以及其他的因素。

⊖ 术语抖动（*jitter*）是网络行话，其值的数学术语是标准差。

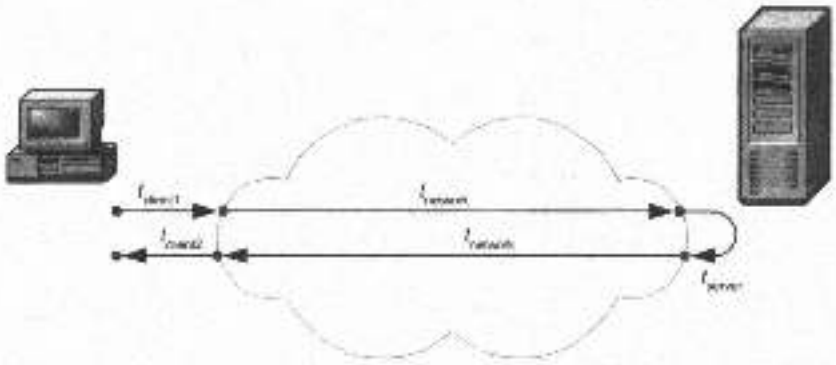


图6.4 网络响应时间和往返时间

- **往返时间**（round trip time, RTT）。这是从源节点到目的节点，然后再返回源节点的数据传输的净时间。它不包括目的节点生成回复的时间。这意味着：

$$RTT = t_{network} + t_{network} \tag{6.4}$$

与网络响应时间不同，RTT允许单独测量网络运输的能力，也正因为这样，允许我们来提高这一能力。而网络响应时间总是从一个够体上刻划网络。

如果每一个方向上的传输时间不同，RTT就是一个很有用的特性。类似于单向的延迟，RTT可以通过平均值和最大值（预定义概率下）来测量。

根据应用的类型，我们可以使用一套特定的延迟特性。例如，考虑在互特网上的音乐广播。因为这一服务不是交互性的，它对单个分组允许较大的延迟，有时候可以是几分钟。但是，延迟变动不能超过100ms到150ms。不然，播放的质量就会显著地降低。因此，在这种情况下，对网络的需求必须包括对平均延迟变动的限制，或对延迟变动的最大值的限制。

6.3.3 信息率的特性

信息率（information rate）总是以某一时间区间来测量，它是传输数据总量除以时间长度的结果。这表明信息率总是表示一个平均值。然而，根据测量的间隔长度，这一特性有不同的名字。

持续信息率（sustained information rate, SIR）被定义为一个相对较长的时间段，长到足够让我们讨论信息率的稳定行为。SLA中需要指定这一值的监测时间段（例如，10秒），这意味着，每过十秒钟，我们就测量一次信息流的速率，并且将它和要求相比较。如果这一控制措施没有被执行的话，当用户和服务提供商意见不一致的时候，用户就无法争取他们的权利。假设在一个月的某一天，提供商没有传输用户流量，但在所有其他的日子里，允许用户超过预定义的额度，这样，每个月的平均速率仍在约定范围内。在这样的条件下，只有有规律地控制信息率才能维护用户的权利。SIR是一个中期特性，SIR是**承约信息率**（commitment information rate）的同义词。

最高信息率（peak information rate, PIR）是在约定的较短时间段T内，用户流量允许达到的最大速率。

这一时间段通常称为**突发时间段**（burst period）。显然，在传输流量时，只有在一个确定程度的概率下，才能谈论这一值。例如，这一特性的要求可以用如下的方式形式地表述：在0.95的概率下，信息率在10ms内不能超过2Mb/s。通常，这一概率会被忽略，因为假设它接近于1。PIR是一个短期的特性。

PIR可以用来测量网络承载那些突发的、造成网络拥塞的负载的能力。如果SLA约定了SIR和PIR，那么突发时间段必须伴随着相对平静的时间段，在平静的时间段内，速率降到平均值之下。

如果不是这样，平均信息率将不会等于约定的值。

突发尺寸（burst size）（通常用B表示）用来衡量在拥塞时，临时存储数据所要求的交换机的缓存容量。突发尺寸等于在允许的最高负载时间段内，到达交换机的总数据量。

$$B = \mathrm{PIR} \times T \tag{6.5}$$

在第3章中，我们提到了流量速率的另一个参数——**流量突发系数**（traffic burst coefficient），有时称为**突发性**（burstiness）。我们将这一系数定义为较长时间段内流量平均速率和某些较短时间段内最大速率的比值。时间段的不确定性使突发系数成为流量的定性特性（*qualitative characteristic*）。

数据传输速率可以在任何两个网络接口之间测量：在客户计算机和服务器之间，在路由器的输入和输出端口之间等。为了分析和调试网络，了解特定网络元素的带宽非常有用。需要指出的是，由于在不同网络元素间数据传输的连续性，任何混合的网络路径的带宽等于具有最小带宽的网络元素的带宽。因此，最大数据传输速率总是由具有最小带宽的元素所限制。为了增加混合的路由的带宽，我们需要首先关注最慢的元素，也就是**瓶颈**（bottleneck）。

6.4　可靠性

为了描述服务可靠性，常常会使用以下两个特性：

- 总分组流中丢失分组的百分比
- 服务可用性

这两种特性都从用户的角度描述了运输服务的可靠性。它们之间的区别在于，它们在不同的时间范围内概括了可靠性。第一种特性是短期的；第二种特性是中期或长期的。

6.4.1　分组丢失特性

这一特性被定义为丢失的分组数量和传输分组的总数量之比：

$$\text{分组丢失率（Loss packets ratio）} = N_L/N \tag{6.6}$$

这里，N等于在某段时间段内，传输的分组总数，N_L等于在这段时间段内丢失的分组数。

6.4.2　可用性和容错

一些可靠性的特性可以用来描述单个设备的可靠性，例如平均失效间隔时间（*mean time between failure, MTBF*）、失效概率（*failure probability*）、失效率（*failure rate*）。但是，这些特性只适用于衡量简单的元素或设备的可靠性，这些简单设备中任何部件的失效都会导致整个设备无法使用。对于由许多部件组成的复杂系统来说，即使它们中有个别部件失效，整个系统还是可用的。

在这一关系中，另有一套特性可以用来衡量复杂系统的可靠性。

可用性（availability）是系统或服务的可用时间与系统总时间之比。可用性是一个长期的统计特性。典型的测量时间间隔是天、月或年。电话网络的通信设备是高可用性系统的一个例子，因为最好的样本具有五九（*five nines*）的可用性。这意味着，这一设备的可用性是99.999%，对应于每年的失效时间略多于五分钟。数据网络的设备和服务只能争取达到这一高可用性。然而，现在已经达到了三九（*three nines*）的可用性。

服务可用性是一个通用的特性，终端用户和服务提供商都会用到。

另一个用来衡量复杂系统可用性的特性是**容错**（fault tolerance）。这是系统屏蔽单一系统部件故障的能力。

例如，如果交换机配备有两个平行工作的交换结构，那么，一个交换结构失效并不会造成整

个交换机的失效，但是，这一交换机的性能将降低，因为和原来相比，它需要花两倍的时间处理分组。在容错系统中，一个元素的失效将导致性能的降低，而不是整个系统的失效。另一个容错实现的例子是使用两条物理链路连接交换机。在正常工作模式下，流量以2C Mb/s的速率在两条链路上传输。如果一条链路失效了，容错系统将继续以C Mb/s的速率传输流量。但是，由于很难得到系统或服务性能降低的定量衡量，容错常被用作定性特性。

之后，我们将考虑保障云输服务高可靠性的最常用方法。

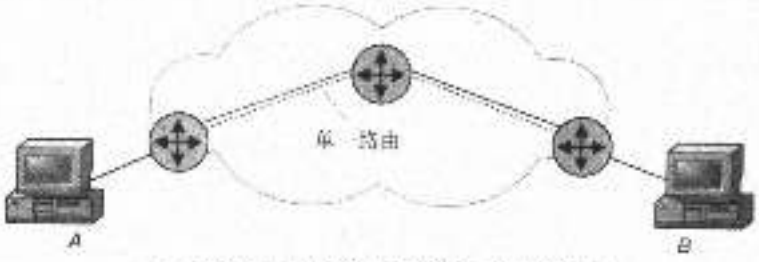
6.4.3 可替换的路由

服务可用性可以通过两种方法提高：

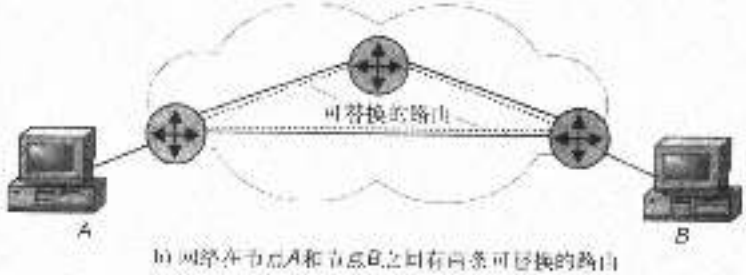
- 第一种方法使用很少失效的可靠的网络元素（*reliable network element*）。但是，这一方法总是被电子部件制造流程的技术所限制（集成电路、印制电路板等）。
- 第二种方法也很著名。它基于将冗余（*redundancy*）引入到系统的设计中：系统的关键元素必须以多个重复的实例存在。这样，如果一个元素失效了，重复性将保障系统的运作。因此，在网络主干上工作的交换机和路由器必须配备有冗余的部件——电源、处理器和接口。

为了保障传输服务所要求的可用性程度，网络需要具有冗余。达到这一目的的主要方法是可替换的路由（alternative route）。图6-5a是一个网络的例子，它没有在节点*A*和节点*B*之间采用用于流量传输的可替换路由。因此，这一网络设计没有提供容错。服务提供商必须依靠网络元素的容错——从节点*A*到节点*B*的路由上的数据链路和交换机。

图6-5b显示了一个可以用两条路由传输节点*A*和节点*B*之间流量的网络。当一条路由上的设备失效时，第二条路由依然可用，网络继续向客户提供服务。当然，从一条路由到另一条路由的转换需要一些时间。在某些转换的时间段中，用户流量可能丢失。因此，减小这一时间段是保障网络容错技术的主要目标之一。



a) 网络在节点*A*和节点*B*之间没有可替换的路由



b) 网络在节点*A*和节点*B*之间有两条可替换的路由

图6-5 可替换的路由

网络中有几种使用可替换路由的方法：

- 只有当主路由失效后，网络才确定可替换路由。这意味着，对每一个信息流来说，网络交换机的转发表中只指定了一条路由。当一个流所在路由上的通信链路或交换机失效后，网络交换机使用特殊的路由协议来寻找可替换的路由。通常，这会花去几十秒或几分钟的时间。

取决于网络的规模和拓扑的复杂性。这是使用可替换路由的最慢的方法。在转换时间段中，用户数据的丢失不可避免。

- 网络事先找到两条路由，并且同时使用它们，这样就创建出了用户看不见的冗余流。在网络输出端，只会选择一条流，其中的数据被传送给用户。这样，其中的一条路由被认为是主要路由，另一条是备用路由。当主路由失效时，用户通过备用路由接收数据。这一方法是最快的，因为它保障了用户流的最高QoS。但是，这也造成了网络性能的显著降低，因为网络传输了两个流，而不是一个。通常，这一方法用来对一小部分非常重要的数据流提供服务，它们需要高度的服务可用性。
- 网络事先找到两条路由，但是只使用其中一条。当主路由失效时，向可替换路由转换的过程比第一种方法快得多，因为系统不需要花时间寻找可替换路由。这一方法也比第二种方法有更高的经济性。但是，数据丢失比第二种方法多，因为已经送往失效路由的数据丢失了。路由上的第一个交换机需要一些时间才能知道网络中出现失效，主路由不再有效。

计算机网络主要使用第一种方法和第三种方法进行路由替换。基于第二种原理（两条活动路由）的技术只用在某些计算机网络中，它们必须保障非常高的可靠性。第二种方法广泛使用在高速传输网络中，它们为电话和计算机流量创建了可靠链路基础结构。

6.4.4 数据重传和滑动窗口

当其他保证可靠性的方法失效、分组丢失时，分组重传的方法会被用到。这些方法要求使用基于连接的协议。

为了确保确实需要重传数据，发送端对发送分组编号；对于每一个分组，发送端希望从接收端收到**肯定确认**（positive acknowledgement，ACK）。ACK是一种特殊的分组，或是数据分组中一个特殊的域，它通知发送端源分组已经被收到，数据是正确的。为了组织这样的编号，需要建立逻辑连接的过程，因为它提供了参考点，从那里开始编号。等待确认的时间是有限的——发送每个分组时，发送端启动一个计时器，如果预定义的时间段已过去，但还没有收到确认的话，发送端认为分组已丢失。如果目的节点接收到数据损毁的分组，它可能发送一个**否定确认**（negative acknowledge，NACK），指出分组需要重传。

有两种方法来组织交换确认的过程：停等方法和滑动窗口方法。

停等方法（the idle-source method）要求发送分组的源发送端等待接收端的确认（或是肯定确认或是否定确认）。只有在收到肯定确认[⊖]之后，发送端才可以发送下一个分组。如果时间段过了后还没有收到确认，或是接收到一个否定确认，认为分组已经丢失，需要重传。图6-6a显示出，这极大地降低了数据交换的性能。换句话说，链路使用率非常低。虽然发送端有能力在发送完前一个分组后立即发送下一个，但发送端需要等待确认。这一差错纠正的方法所造成的性能降低在低速链路上尤其明显（即广域网链路）。

第二种差错纠正的方法称为**滑动窗口方法**（the sliding window method）。为了增加数据速率，这一方法允许发送端以连续的方式（即以最大的速率，而不需要接收这些分组的肯定确认）传送特定数量的分组。这种方法中，可以传送的分组数称为**窗口大小**（window size）。图6-6b显示了窗口具有W个分组的这一方法。

在开始的时刻，还没有分组被发送出去，窗口定义了分组的范围，从数字1到W，1和W包含在内。源发送端开始传输分组和接收确认。为了简化起见，我们假设确认以和它们对应的分组发送顺序相同的顺序到达。在t_1时刻，接收到第一个确认（ACK1）后，窗口移动了一个位置，定义了一个新的范围：从2到（$W+1$）。

⊖ 为了简化起见，我们以后将用确认（acknowledgement）表示肯定确认（positive acknowledgement）。

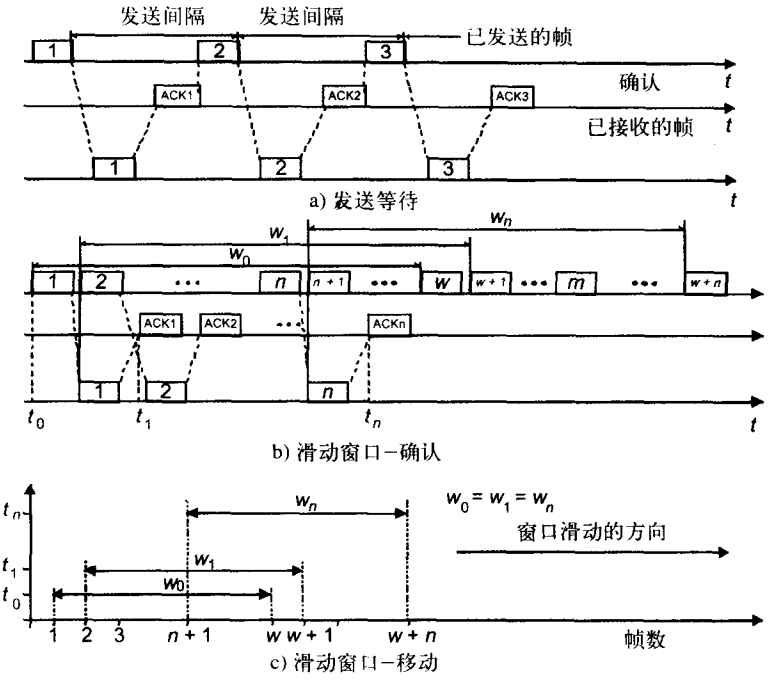


图6-6 恢复损毁或丢失分组的方法

发送分组和接收确认的过程互相独立。考虑在一个任意时刻，t_n，源发送端接收到了标号为1到n的所有分组的确认。窗口向右滑动，定义了一个新的允许传输的分组的范围。这一范围从$(n+1)$到$(W+n)$。源发送端中的所有分组可以被分为如下几类（参见图6-16b）：

- 标号为1到n的分组已经被发送，这些分组的确认已经被接收，这意味着它们在窗口左边界的外侧。
- 标号从$(n+1)$到$(W+n)$的分组位于窗口的范围内。这些分组可以被发送，而不需要等待确认。这一范围又可以被进一步划分为以下区间：
  - 标号从$(n+1)$到m的分组，它们已经被发送，但是它们的确认还没有被收到
  - 标号从m到$(W+n)$的分组，它们可以被发送，但还没有发送
- 标号大于$(W+n)$的所有分组位于窗口右边界的外部，因此，还不能被发送。

图6-6c描述了沿着分组标号的顺序滑动窗口的过程。这里，t_0是起始时刻，t_1和t_n分别是第一个分组确认和第n个分组确认到达的时刻。每次一个确认到达，窗口向右滑动，它的大小始终保持不变，等于W。

这样，当传输标号为n的分组时，超时时间在源节点处设定。如果在这段时间内，分组n的确认没有到达，分组n被认为丢失，必须被重传。

如果确认的流有规律地到达，那么源方总是有一些有权发送的分组，交换速率达到了特定链路可能的最大值。从滑动窗口方法的描述中，很显然发送等待是这一算法的一个特例，其窗口大小为1。

滑动窗口算法的某些实现不需要接收端对正确接收的分组发送确认。如果接收到的分组没有间隙，接收端只要对最后接收的分组发送确认，这一确认被发送端理解为，所有之前的分组都被递送成功。

有些方法使用了否定确认。否定确认有两种类型——组确认和选择确认。组确认包含了一个分组编号，从这一分组编号开始，我们需要重新传输源方发送的所有分组。选择否定确认要求重新

传输单个分组。

滑动窗口方法有两个可以显著影响发送端和接收端数据交换效率的参数。它们是窗口大小和发送端等待确认的超时。超时的选择取决于分组的延迟。

在可靠的网路中，分组很少丢失或损毁，常常需要增加窗口大小，以增加数据交换速率。使用这一方法时，发送端以很小的停顿发送分组。在不可靠的网络中，窗口大小必须减少，因为在分组频繁丢失、或损毁、或同时发生的情况下，重传的分组数快速上升。如果发生这一情况，网络的带宽得不到有效的利用，有效网络带宽就降低了。

窗口大小可以是这一算法的一个固定参数，这意味着这一参数在连接建立时选定，在整个过程中不会被改变。

这一算法也有适应性（adaptive）的版本，其中，窗口大小根据网络的可靠性和负载，在整个过程中会发生变化。在低网络可靠性和高负载的情况下，发送端减小窗口大小，试图找到一个最优的数据传输模式。这一算法中网络可靠性由一些分组丢失的征兆来决定，诸如肯定确认超过时限，或是到达了多个针对某一分组的重复确认。多个重复确认表示目的节点中，下一分组已经超过了时间限制，它要求重传这一分组。

目的节点也可以改变窗口大小。这可能发生在目的节点过载、无法及时处理所有到达分组的时候，我们将在学习消除网络拥塞的问题时，在第7章的7.6节考虑这一问题。

也有些滑动窗口算法使用了字节数作为窗口大小，而不是分组数。TCP协议是这一方法中最著名的例子。

滑动窗口方法比停等方法更难实现，因为发送端必须将所有肯定确认还没有收到的分组存储在缓存中。此外，它还需要追踪这一算法的多个参数，包括窗口大小、已收到确认的分组数、在接收到新确认前可以发送的分组数。

6.5 安全

计算机网络是访问多种信息的杰出工具，也是极佳的通信工具。但是，计算机网络也有其不好的一面。计算机网络不好的地方在于，它将你托付给网络的信息的完整性和保密性置于潜在威胁中。例如，具有与因特网持续连接的公司的信息资源常常受到入侵者的攻击。通过拨号连接到因特网的用户也会暴露于攻击之下。他们计算机上存储的信息可能遭受邮件蠕虫或即时通信系统（例如，ICQ）漏洞的攻击。

一些统计数据

在《Issues and Trends》的年度报告中指出：2002年CSI/FBI计算机犯罪和安全调查于2002年4月出版，计算机犯罪显著地增加。大约90%的被调查者（主要是大公司和政府机构的雇员）报告说，在过去的12个月中，他们的组织机构中发生了安全事件。大约80%的被调查者认为这些安全事件造成了经济损失。44%的被调查者定量地评估了损失。根据他们报告的数据，总经济损失超过了4亿5 500万美金。

入侵者可以使用因特网和公司网试图非授权地访问或破坏计算机信息。没有人可以保证某些对现状不满的雇员不会不当使用特权，试图访问他们没有权利阅读的文件。试图破坏信息的行为（例如，删除文件、造成计算机无法工作）也有可能会发生。

显然，网络用户希望保护他们的信息免受这类攻击。网络用户信息的安全水平是另一个重要的网络特性。安全水平不是定量的特性。它只能定性地衡量，例如，高水平、中水平和低水平。通常，为了恰当地衡量网络安全程度，我们需要咨询专家。

6.5.1 计算机和网络安全

不同的信息保护工具可以被分为两类：

- **计算机安全工具**（computer security tool）是为了保护位于局域网中或单个终端用户计算机上的内部信息资源。
- **网络安全工具**（network security tool）是为了保护在网络上传输的信息。

这两类工具中的安全功能差异显著。在第一类中，需要保护内部局域网中的所有资源免受非授权访问，包括硬件（服务器、磁盘阵列、路由器）、软件（操作系统、数据库管理系统、邮件服务等），以及文件中存储的数据和RAM中处理的数据。为了实现这一目标，我们需要检查所有从公共网络（目前，因特网是最主要的公共网络）到达本地网络的流量，并且设法阻止某些可能帮助入侵者非法使用保密信息资源的外部访问。

这类保护工具中最常用的是**防火墙**（firewall），它安装在内部网和因特网之间的所有连接上。防火墙是互联网络的过滤器，它检查协议所有层的报文交换，不允许可疑的信息进入受保护的网络。

防火墙也可以在网络中使用，以保护各个子网。具有多个独立部门的大公司可能需要这一配置。这类问题也可以由具有内置安全工具的操作系统和应用程序（例如数据库管理系统）、以及内置安全功能的硬件来解决。

在第二类工具中，在网络边界外的信息需要被保护。通常，这些信息以IP分组的形式在服务提供商的网络上传输。目前，因特网被大部分的公司所使用，它不但被作为一个分布在无数网站上的强大的信息源，也是一个相对便宜的运输环境，允许总部网络和其他部门的网络连接。它也可以连接众多移动用户和远程工作者。在大多数情况下，保障因特网上传输的信息不被损毁、破坏、或者被未授权的第三方查看显得非常重要。目前，出于这一目的，**虚拟专用网络**（virtual private network，VPN）是最常用的工具。

单个计算机可以通过使用不同的工具，有效地防止外部入侵。例如，可以简单地锁住键盘，或卸下硬盘放在安全的地方。加入网络的计算机不能完全和外部世界隔绝，因为他们必须和其他计算机通信，有些计算机之间可能相距甚远。因此，与保障单个机器的安全性相比，保障网络的安全性是一个更复杂的任务。如果你的计算机连在网络上，远端用户和你的计算机建立逻辑连接的情形是很常见的。在这种情况下，保障安全意味着使这些连接处于可控状态——对于每台本地计算机上存储的信息、对于外部设备的访问权限、对于在每台网络计算机上进行特定管理操作的权限，每一位网络用户都必须具有严格定义的访问权限。

除了远程登录网络计算机可能造成的问题外，网络也暴露在其他的威胁下：网络上传输的报文的探测和分析，以及生成的伪造流量。大多数的网络安全工具是为了防止这类安全事件。

现在，当大部分公司在创建企业网时，由租借线路转向公共网络（因特网、帧中继），网络安全领域就变得特别重要。

6.5.2 数据保密性、完整性和可用性

首先，一个安全的信息系统保护数据免受未授权访问；其次，它总是能提供所需要的数据给授权用户；最后，它可靠地存储着信息，并保证它们的不变性。这样，一个可靠的系统具有保密性、可用性和完整性。

- **保密性**（confidentiality）保证数据只能被那些具有访问这些信息权限的用户获得。他们被称为授权用户。
- **信息可用性**（information availability）保证授权用户总是可以访问所需要的信息。
- **完整性**（integrity）保证数据将维持正确的值。这通过阻止未授权用户的访问，防止他们改

变、修改、删除、创建新的数据来保障。

入侵者可能试图违反信息安全的所有组成部分——可用性、完整性、保密性。根据系统的用途、数据的类型、可能遭受的威胁，安全的要求可能会改变。很难想像人们会认为一个系统的完整性和可用性不重要。但是，保密性并不总是必须的。例如，如果你在因特网的网站上发布信息，希望它能被最广泛的用户使用，这时，保密性并不需要。但是，完整性和可用性的要求依然非常重要。

如果你不采取特殊的措施来保障数据完整性，入侵者可能改变存储在服务器上的数据，对你的公司造成损害。例如，恶意的用户可能在Web服务器上发布的报价单中进行一些这类改动，这将降低公司的竞争力。他们也可能破坏公司提供的免费软件的代码，这将对公司的声誉造成损害。

在这个例子中，保障数据的可用性也很重要。在投入了大量的资金创建并支持网站后，公司有理由希望这一投资能以某些方式带来回报，例如客户数量增加、销售增加等。但是，入侵者可能发起一次攻击，造成服务器上发布的数据不能被需要这些信息的用户获得。这类攻击的例子包括用包含错误返回地址的IP分组淹没服务器。在这一协议的内部逻辑中，这些分组会造成超时，最终使服务器无法处理所有的用户请求。这一攻击是拒绝服务（DoS）攻击的一个特例。

注意　保密性、可用性、完整性的概念不但可以与信息相关，还可以与其他数据网络上的资源相关，包括外部设备和应用程序。例如，不受限制地访问打印机可能使入侵者得到被打印的文档，改变打印机的设置，甚至造成设备失效。就打印机而言，保密性可以这样理解：只有那些授权使用这一设备的用户才有权力访问它。此外，即使是授权用户，他们也只能执行他们允许的操作。这一例子中的可用性是指，设备必须在任何时候都可用。就完整性而言，它可以被定义为对特定设备的设置不能改变。

6.5.3　网络安全服务

用于数据保护的不同软件和硬件通常使用相似的方法、技术和技术解决方案。让我们来考虑最重要的几种。

- **加密**（encryption）是所有信息安全服务的基础，无论是对于认证或授权系统、对于创建受保护信道的工具、还是对于保障数据存储安全的方法。加密的过程将信息从可读的形式（明文）转化成加密的数据（密文）。任何加密的过程必须伴随一个解密的过程。这一解密过程应用于密文，将它转变回可读的形式。一对这样的过程（加密和解密）称为**密码系统**（cryptographic system）。
- **认证**（authentication）防止对网络的未授权访问，只允许合法的用户登录。术语认证（*authentication*）起源于拉丁语，意思是验证真实性。需要认证的对象，除了用户外，还包括不同的设备、应用、信息。例如，向公司服务器发送请求的用户必须提供他们的身份，确保他们在和公司的服务器通信。换句话说，客户和服务器都必须经过互相认证的过程。这是应用层的认证。在两个设备间建立通信时，就可能需要更低的、数据链路层的认证过程。数据认证指证明数据的完整性，以及数据确实是从声称提供这一信息的用户那里接收。为达到这一目的，**数字签名**（digital signature）方法被广泛使用。认证不应该和身份识别混淆起来。
- **身份识别**（identification）指用户向系统表明个人的身份识别符；认证是验证用户是否确实是他/她所声称的人的过程。特别是，当用户登录后，用户必须提供密码，来证明用户确实是这个特定的标识符所属于的那个人。用户的标识符和其他对象（例如，文件、进程、数据结构）的标识符以同样的方式使用。他们并不直接和安全相关。
- **授权**（authorization）是控制已经认证的用户对系统资源访问的过程。授权系统提供给每个用户一定的权力，这些权力由管理员赋予。除了提供用户访问文件、目录、打印机等的权力

外，授权系统可能控制用户的权限（即执行特定操作的能力），例如，本地访问服务器、设置系统时间、创建数据的备份、关闭服务器。被授予特定访问权力和权限的、经认证的用户称为授权用户。

- **审计**（auditing）是将所有与访问受保护系统资源相关的事件登记入系统日志的过程。当代操作系统的审计子系统允许我们通过方便的图形用户接口，区分系统管理员感兴趣的不同事件。审计和监测工具可以检测和登记重要的安全相关事件、试图创建新系统资源的事件、访问、修改或删除已有资源的事件。审计用来检测入侵企图，甚至是失败的入侵企图。
- **受保护信道技术**（protected channel technology）用来保障公共网络上（例如，因特网）数据传输的安全性。受保护的信道意味着遵守如下三个要求：
  - 在建立连接时，用户之间互相认证，这可以通过诸如交换密码来实现。
  - 通过使用受保护的信道防止未授权访问，来保护传输的报文，例如，通过数据加密。
  - 保障通过受保护信道达到的报文的完整性，这可以通过诸如同时传输数字签名来实现。

在创建VPN时，受保护信道技术得到了广泛的使用。

6.6 仅用于服务提供商的特性

让我们来了解一下服务提供商在衡量它们网络的效率时，会使用到的主要特性。这些特性通常是定性的。

6.6.1 可扩展性和可延拓性

术语可扩展性和可延拓性有时候被用作同义词。这是不正确的，因为这两个术语都有严格定义的、独立的意思。

- **可扩展性**（extensibility）是相对容易地增加用户和新网络部件（例如，计算机、交换机、路由器、服务）、增加网络段电缆长度、将现有设备替换为新设备（更高级更强大的设备）的可能性。扩展网络的便捷性有时候只能在一定程度上得到保障。例如，基于单一段细同轴电缆的以太网具有很好的可扩展性，因为它可以便捷地连接新的工作站。但是，这样的网络受到连接工作站的数量的限制，通常这一数量不能超过40。虽然这一网络允许更多数量的工作站物理连接到段上（可达到100），但是，这样的话，网络性能将显著地下降。这一限制表现出了网络较差的可扩展性，虽然这一系统的可延拓性相当好。
- **可延拓性**（scalability）指大量增加网络节点数量和链路长度，但不降低网络性能的可能性。为了保障网络的可延拓性，我们需要使用额外的通信设备，并遵守特定的规则来建造网络。通常，可延拓的解决方案具有多层的层次结构，这一结构可以在不改变项目的主要思想的条件下，在每一层增加新的网络元素。因特网是可延拓网络的一个例子，因为它的技术（TCP/IP）能在世界范围的规模上支持网络。因特网的组织结构已经在第5章中做了介绍，它由多个层次组成：用户网络、本地ISP网络等，一直到国际ISP网络。整个因特网所基于的TCP/IP技术也可以建造层级网络。主要的因特网协议——IP——基于两层的模型。低层由单个网络创建（通常是公司网），高层是将这些网络连接在一起的互联网络。在TCP/IP栈中，有一个概念叫做自治系统。自治系统包括所有的单一ISP的互联网络，所以自治系统是一个较高的层次。因特网上的自治系统将解决方案简化为合理的路由问题。首先找到自治系统之间合理的路由，然后，对于每个自治系统，在它的边界内找到合理的路由。

为了实现一个可延拓的解决方案，不但需要网络本身是可延拓的，主干设备也必须是可延拓的，因为网络的增长不能导致设备需要不时地更新。因此，主干交换机和路由器通常基于模块原理，这使得它们可以很容易地增加接口数量和分组处理性能。

6.6.2 可管理性

网络可管理性（network manageability）是一个定性的特性，表示了网络集中式控制主要网络元素状态、检测和解决网络问题、分析性能、规划网络增长的能力。可管理性意味着网络中必须包括自动化的管理和控制工具。这些自动化工具使用通信协议和网络软硬件交互。

理想情况下，**网络管理系统**（network management system，NMS）监测、控制、管理每一个网络元素——从最简单到最复杂的设备。同时，这一系统将网络当作一个整体，而不是一系列单个的设备。

一个好的NMS监测着网络，一旦发现问题，它就采取特定的行动，修正这一问题，并且通知管理员所发生的事件和采取的行动。同时，NMS必须积累数据，基于这些数据可以规划进一步的网络开发。最后，NMS需要提供便捷的用户接口，允许所有的操作在单一的控制台上完成。

NMS的用处在大规模网络中变得尤其明显，例如公司网的公共广域网。如果没有NMS，这样的网络需要由专业化的维护支持人员长期驻扎在网络设备安装的每个城市、每幢建筑。因此，大量的支持人员就变得必不可少。

在网络管理系统领域还有很多未解决的问题。人们需要便捷的、精简的、多协议的网络管理工具。大多数现有的工具实际上不是网络管理工具，它们只执行网络监测和报告重要事件（例如设备失效）的任务。

6.6.3 兼容性

兼容性（compatibility），或称为**整合能力**（integration capability），意味着网络具有包含多种软件和硬件的能力（即支持不同协议栈的不同操作系统、不同厂商的不同软件硬件产品可以共存）。由这些不同元素构成的网络称为**异构网络**（heterogeneous network）。如果异构网络运作良好，它就被称为**集成网络**（integrated network）。建造集成网络的主要方法是使用根据开放标准和规范设计的网络模块。

小结

- 对计算机网络的主要要求是保障较高的服务质量（QoS）。使用这一术语的广义含义时，QoS的概念包括了所有可能的网络特性和用户所希望的服务。
- 对网络服务质量的要求通过使用标准化的特性来表达。
- 运输服务的质量通过使用以下几组特性来衡量：
  - 性能
  - 可靠性
  - 安全性
  - 仅用于服务提供商的特性，这包括可扩展性、可延拓性、可管理性和兼容性
- 网络性能通过使用以下两类统计特性来衡量：*信息率特性*和*传输延迟特性*。第一组包含了在突发时间内的持续速率和最大速率，以及突发时间的长度。第二组包含了平均延迟的值、延迟变动的平均值（抖动）、变动率、延迟和延迟变动的最大值。
- 网络可靠性使用多种特性来衡量，包括分组丢失的百分比、可用性系数（指系统可用的时间）、容错（系统在某些部件失效的情况下继续工作的能力）。
- 网络提供的运输服务的可靠性由网络部件的可靠性（通信链路和通信设备）、可替换路由的可用性、对丢失或损毁分组的重传来保障。
- 网络安全工具包括：
  - 用来保护位于局域网内单个计算机上的内部信息资源的计算机安全工具
  - 用来保护信息在网络上传输的网络安全工具

- 信息安全的主要特性有：
  - 保密性——保证数据只可以由具有访问这些信息权力的授权用户获得
  - 可用性——保证授权用户总是可以访问数据
  - 完整性——保证数据维持正确的值，这通过防止未授权用户访问、修改、删除、创建新信息来实现
- 为了保护网络信息，加密、认证、授权、审计的机制被使用。网络上的数据传输通过使用受保护信道技术来实现。

复习题

1. 特性和需求的区别是什么？
2. 从广义上说，服务质量（QoS）的概念包含了哪些特性？
3. 哪些QoS特性只有终端用户感兴趣？哪些只有服务提供商感兴趣？哪些服务提供商和终端用户都感兴趣？
4. 从狭义意义上说，什么是QoS特性？
5. 哪些性能特性只有服务提供商感兴趣？
6. 服务水平约定由哪几方订立？
7. 假如你需要在网络上传输IP电话应用的流量，请列出一些你需要在SLA中包括的特性。
8. 测量分组延迟的结果应该使用什么类型的信息表示？
9. 与抖动相比，使用变动率这一特性有什么好处？
10. 在定义往返时间时没有考虑哪一部分的影响？
11. 是否可能在传输数据时发生很大的延迟，但没有抖动？
12. 请列出突发（burst）的参数。这些参数是否是独立的？
13. 平均流的速率是否取决于分组延迟？
14. 短期范围内使用了运输服务可靠性的哪些特性？中期范围内使用了运输服务可靠性的哪些特性？
15. 请描述保障网络可靠性的两种主要方法。
16. 使用可替代的路由有几种方法可以用来增加流量传输的可靠性？它们有什么优点和缺点？
17. 信息安全有哪两个部分？
18. 可扩展性和可延拓性有什么区别？

练习题

1. 两个交换机由两条物理链路连接，以增加可靠性（图6-7）。对于两种使用可替代路由的方法，请估计链路失效时丢失的数据量。这两种可替代路由的方法是：方法2，“网络事前找到两条路由，并使用它们”；方法3，“网络事先找到两条路由，但只使用一条”。每条链路的长度是5 000km，数据传输速率是155Mb/s，链路中信号传播速度是200 000km/s。在两种情况下，交换机S2检测到了链路失效，并在10ms内切换到备用链路。

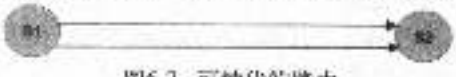


图6-7 可替代的路由

2. 如果数据使用基于停等算法的协议传输，请估计链路利用率。传输速率为100Mb/s，往返时间（RTT）为10ms，分组不会丢失或损毁。分组大小是固定的，等于1 500字节。确认的大小可以被忽略。
3. 如果使用链路传输分组而不让源方空闲等，请计算为达到这一要求，窗口的最小尺寸为多少？传输速率为100Mb/s，RTT为10ms，分组不会丢失或损毁。分组大小是固定的，等于1 500字节，确认的大小可以被忽略。

第7章 保证服务质量的方法

7.1 引言

目前，服务质量的方法是分组交换网络技术中最重要的技术之一，当代多媒体应用（例如IP电话、视频和电台广播、互动远程学习）的运行离不开它们的实现。这些方法影响如下三组的网络特性：

- 信息率
- 分组延迟
- 分组丢失

这些特性的定义在上一章中已经做了介绍。

QoS关注流量传输时通信设备的队列所造成的影响。QoS方法使用了多种队列管理、预留和反馈的算法，这些方法将负面影响减小到某一最小值，让用户可以接受。

队列是分组交换网络的一般属性。分组交换原理本身就假设了分组交换机的每个输入输出接口具有缓存。在网络拥塞时进行分组缓存是支持突发性流量的主要机制，这一机制保障了这类网络较高的性能。另一方面，队列意味着网络上分组传输具有不确定的、可变的延迟，这是对延迟敏感流量造成问题的主要原因。由于分组网络运营商对传输这类流量非常感兴趣，它们要求有一些特殊的工具，在最大化网络负载和满足所有类型网络流量所要求的QoS之间，实现一种折衷。

所有这些特性都描述了队列的负面影响。事实上，如果发生了网络拥塞，通常会在拥塞期间减小流的速率，造成分组延迟，甚至分组丢失。如果队列完全占满缓存，分组丢失就会发生。

QoS方法使用多种机制来减小队列的负面影响。这些机制的范围相当广泛，我们将详细地介绍它们。它们中的大多数考虑了网络中现有的多种类型的流量。

流量工程的方法可以对QoS方法进行补充，它管理流量路由，以平衡流量、消除队列溢出。

7.2 应用与QoS

7.2.1 不同类型应用的QoS要求

第1章中所介绍的当前不同类型网络融合的趋势，意味着数据网络现在必须负担各种类型的流量，不仅仅是文件访问和电子邮件信息。

在前一小节中，我们列出了用来衡量流量传输质量的不同QoS特性。当网络同时传输不同类型的流量时（例如，Web应用的流量和语音流量），这些特性尤其重要。这是因为不同类型的流量对QoS特性有不同的要求。要同时满足所有类型流量的所有QoS要求，这是一项非常困难的任务。因此，我们常常使用以下的方法：将网络上现有的所有类型的流量划分成多个流量类型，然后，设法同时满足每一个类型对QoS的要求。

人们已经做了许多优秀的研究工作，试图根据应用所产生的流量，对应用进行分类。以下三种应用特性是主要的流量评判标准：

- 应用所生成流量的信息率的相对可预测性
- 应用对分组延迟的敏感性
- 应用对分组损毁或丢失的敏感性

7.2.2 信息率的可预测性

就信息率的可预测性而言，所有应用可以分成以下两类：生成流数据的应用、生成突发性数据的应用。

流应用以一个**恒定的比特率**（constant bit rate，CBR）生成数据。如果使用了分组交换，这类应用的流量就是一系列同样大小的分组，每个分组等于B比特，分组之间有同样的时间间隔T（图7-1）。

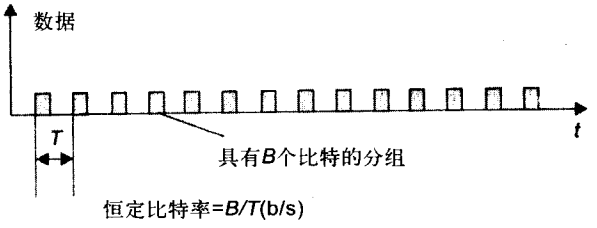


图7-1 流数据

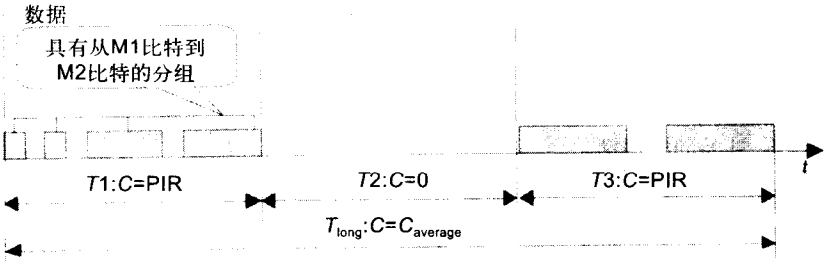


图7-2 突发数据

流数据的恒定速率（CBR）可以通过一段时间T内的平均值来求：

$$\mathrm{CBR} = B/T(\mathrm{b/s}) \tag{7.1}$$

流数据的恒定速率比数据传输协议的额定最大比特率要低，这是因为在分组之间存在停顿。在第12章中我们将看到，使用以太网协议的最大数据传输速率是9.76Mb/s（当帧具有最大长度时），比这一协议的额定速率低（10Mb/s）。

产生突发数据的应用具有低可预测性的特性，因为在一段时间的静默期后，会出现一个突发期，在这一期间，分组一个接一个非常密集地发出。因此，流量具有**可变比特率**（variable bit rate，VBR），如图7-2所示。在文件服务类应用工作时，这类应用所产生的流量密度可能降到零（如果目前没有文件在传输），或是服务器传输文件时，流量密度上升到最大值（只被网络的能力所限制）。

图7-2显示了三个测量的时间段——$T1$、$T2$、$T3$。为了简化计算，我们假设第一和第三个时间段内的最高速率是相等的，等于PIR，这两个时间段有相同的时间长度T。如果我们知道了T的值，就可以计算出突发尺寸B，它等于在突发时间段内传输的比特数：

$$B = \mathrm{PIR} \times T \tag{7.2}$$

这样，$T1$和$T3$时间段的突发尺寸等于B，T2时间段的突发尺寸等于零。

对于这个例子，我们可以计算突发系数。它等于一小段时间内的最高速率和一长段时间内的平均流量速率的比。$T1$和$T3$时间段的最高速率等于B/T，整个测量时间段的平均速率$C_{average} = T1 + T2 + T3$等于$2B/3T$。因此，突发系数等于3/2 = 1.5。

7.2.3　应用对分组延迟的敏感性

应用对分组延迟和变动的敏感性是另一个根据流量类型对应用进行分类的标准。根据对分组延迟的敏感性从低到高的顺序排列，应用的主要类型如下：

- **异步应用**（asynchronous application）。对于这类应用来说，对分组延迟的时间几乎没有限制。在这种情况下，用户在和**弹性流量**（elastic traffic）打交道。这类应用的典型例子是电子邮件。
- **交互应用**（interactive application）。用户可能会注意到延迟，但它们并不对应用的功能造成负面影响。这类应用的一个例子是使用文本编辑器访问远程文件。
- **等时应用**（isochronous application）。这类应用有一个对延迟变动的敏感性的阈值。如果超出这一阈值，应用的功能就会显著降低。例如，在语音传输应用中，如果超出了延迟变动的阈值（100~150msec），语音回放的质量就会显著降低。
- **对延迟过分敏感的应用**（applications oversensitive to delay）。数据传输发生延迟会将应用的功能降为零。这类应用的一个例子是那些实时控制的应用。如果控制信号被延迟，所控制的对象就可能遭到破坏。

通常，应用的交互功能总是会增加它对延迟的敏感性。例如，语音广播可以在分组传输中容忍较大的延迟（但对延迟变动依然很敏感），但是，交互式电话或视频会议就不能容忍延迟。当对话通过卫星来传输时，这一点尤其明显。对话中较长的等待时间常常使参与方感到迷茫，于是，他们会失去耐心，开始同时说话。

注意　上面的分类方法提供了对应用延迟和变动敏感性的细微区分，除了这种分类方法外，另一种分类方法根据同一标准，对应用进行了更粗略划分。根据第二种分类，应用被分成两类——具有弹性流量的异步应用和具有延迟敏感流量（或时间敏感）的同步应用。**异步应用**（asynchronous application）是那些可以容忍多达几秒的数据传递延迟的应用。所有其他那些功能会被数据传递延迟而损害的应用都被归类为**同步应用**（synchronous application）。交互应用既可以被归类为异步应用（例如，文本编辑器），也可以被归类为同步应用（例如，视频会议软件）。

7.2.4　应用对分组丢失的敏感性

最后一种应用分类标准是它们对分组丢失的敏感性。通常，应用属于以下两组：

- **数据丢失敏感的应用**（applications sensitive to loss）。事实上，所有传输字母和数字（文本文件、源代码和数值数组等）数据的应用都会对单个数据段的丢失高度敏感，无论这一段数据有多小。这种丢失往往令所有成功接收的信息变得无法使用。例如，源代码中丢失了一个字节，它就几乎毫无用处。所有传统的网络应用（文件服务、数据库管理系统和电子邮件服务等）都属于这一类别。
- **容忍数据丢失的应用**（applications sensitive to loss）。这一类包含了许多应用，它们传输带有惯性物理过程（inertial physical process）信息的流量。容忍丢失意味着基于正确接收的信息，可以恢复出那一小部分丢失的数据。因此，如果丢失了一个带有多个连续语音度量的分组，在回放时，丢失的数据可以用邻近的值近似地加以替换。这类应用包括大多数与多媒体流量有关的应用（音频和视频应用）。但是，容忍数据丢失的程度是有限的，丢失分组的百分比不能太大（通常，不能超过1%）。

并不是所有的多媒体流量都可以容忍数据丢失。例如，压缩语音和视频对数据丢失非常敏感，因此，它们属于第一类的应用。

7.2.5 应用类别

取决于不同的分类标准（数据速率的相对可预测性、流量对分组延迟的敏感性、流量对分组丢失或损毁的敏感性），一个应用可能属于不同的类。这意味着流应用可以被分类为同步应用或异步应用。同步应用可以对分组丢失敏感，或是可以容忍数据丢失。但是，实践证明，在所有应用特性的组合中，有些组合是目前大多数应用所使用的特性。

例如，具有“生成流量——流、等时、容忍数据丢失”特性的应用对应于诸如IP电话、视频会议、因特网上的语音广播这些流行的应用。另一方面，对于某些特性的组合，很难给出一个现有应用的例子。其中的一个组合是“生成流量——流、异步、对数据丢失敏感”。

对应于特定类型应用程序的固定的特性组合并不是很多。例如，ATM技术最早是为了支持不同类型的流量而开发的，在ATM技术的标准化过程中，定义了四种**应用类别**（application class）：A、B、C、D。对于每种应用类别，它推荐使用一套特定的QoS特性。除此之外，对于不属于这些类别的所有应用，定义了一种特殊的类别（X类），对它来说，应用特性的组合可以是任意的。

目前，ATM类型是最详细、最通用的一种。它不依赖于特定的技术，也不需要我们了解这些技术。表7-1简要地列出了这一分类。

表7-1 流量的类别

流量的类别	特　性
A	恒定的比特率（CBR）
	对延迟敏感
	面向连接
	例子：语音流量、电视流量
	QoS特性：最高信息率、延迟、抖动
B	可变比特率（VBR）
	对延迟敏感
	面向连接
	例子：压缩语音、压缩视频
	QoS特性：最高信息率、突发性、持续信息率、延迟、抖动
C	可变比特率（VBR）
	弹性流量
	面向连接
	例子：终端节点使用基于连接的协议（帧中继、X.25、TCP）时，计算机网络的流量
	QoS特性：最高信息率、突发性、持续信息率
D	可变比特率（VBR）
	弹性流量
	无连接
	例子：终端节点使用无连接的协议（IP/UDP、以太网）时，计算机网络的流量
	QoS特性：没有定义
X	用户定义的流量类型和参数

这一应用分类是当代网络中QoS典型参数要求和机制的基础。

7.3 队列分析

如果你定义好了主要的QoS特性，并且形式化地阐述了对它们的要求，那么你已经解决了问题的一半。用户通过使用一套QoS特性阈值，来系统地阐述对QoS的要求。例如，用户可以指定在0.99的概率下，分组延迟的变动不能超过50msec。

但是，用户如何能保证网络可以成功完成所阐述的这一任务呢？可以采取哪些步骤来保证延迟变动不超过指定的值？用户如何保障输出用户流的平均速率对应于输入用户流的平均速率？

在很长一段时间内，这些问题被认为是无关紧要的。分组交换网络最早是为传输异步流量而设计的。因此，延迟是可以被容忍的。但是，现在，数据网络开始传输不同类型的流量，包括实时流量，QoS方面就变得非常重要了。

为了理解QoS的支持机制，我们首先需要学习网络设备中的排队过程，理解影响队列长度的最重要因素。

7.3.1　M/M/1模型

排队理论是应用数学的一个分支，它研究了排队的过程。我们不打算深究这一理论的数学基础，而只是学习与QoS问题相关的部分结论。

图7-3显示了排队理论的最简单模型，称为M/M/1模型[⊖]。

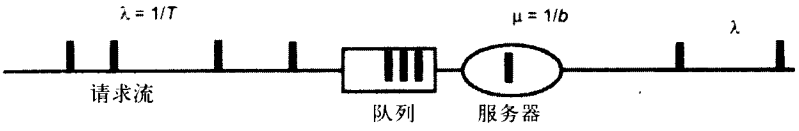


图7-3　M/M/1模型

这一模型的主要元素如下：

- 抽象服务请求的输入流
- 缓存
- 已服务请求的输出流
- 服务器

请求在随机的时刻到达缓存入口。如果新的请求到达时，缓存是空的且服务器空闲，那么请求立即被传递到服务器。服务的时间也是随机的。

如果请求到达时，缓存是空的，但服务器正忙于处理前一个请求，那么到达的请求必须在缓存中等待，直到服务器可用。服务器一旦完成前一个请求的处理，它就被传送到输出口，然后服务器从缓存中取出下一个请求。离开服务器的请求形成输出流。缓存是无限的，这意味着请求不会因为缓存溢出而丢失。

如果新到达的请求发现缓存非空，它就被放进队列中，等候服务。根据请求到达的顺序，依次从队列中提取——也就是说，根据**先进先出**（First In, First Out，FIFO）的服务顺序。

排队理论可以根据输入流和服务时间的特性，估计这一模型的平均队列长度和平均等待时间。

假设两个请求到达之间的平均时间为T。这意味着请求到达的速率（在排队理论中，通常用λ来指代）为：

$$\lambda = 1/T \text{ 每秒请求数} \tag{7.3}$$

这一模型中，请求到达的随机过程由请求到达之间间隔的分布函数来描述。为了得到精简的分析结果，通常假设这些间隔由所谓**马尔可夫**（Markovian，也被称为**普阿松**（Poisson））分布来描述。图7-4显示了这一分布密度。从这一图示中，我们可以很明显地看到输入流的突发性较大，因为请求之间间隔非常小（接近于零）或非常长的概率不为零。间隔的平均偏差也等于T。这样，标准差为$T/T = 1$。

⊖ 这里，1表示这个模型有一台服务器，第一个M代表请求到达间隔的分布函数类型（马尔可夫），第二个M指代服务时间的分布类型（也是马尔可夫）。

假设单个请求的平均服务时间等于b，这表示服务器可以以$1/b=\mu$的速率转发请求到输出口。同样，为了得到精简的分析结果，我们假设服务时间是服从指数分布密度特性的随机变量。

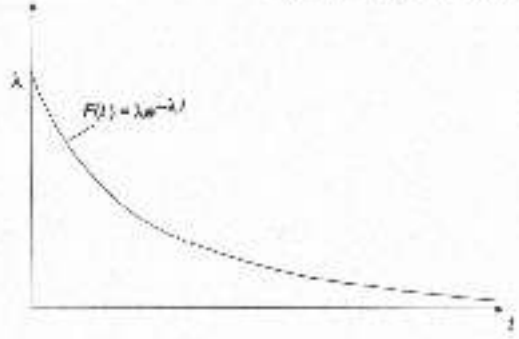


图7-4 输入流的分布密度

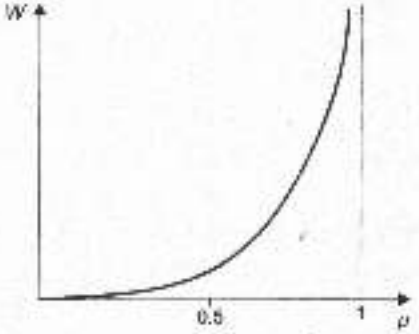


图7-5 平均请求等候时间和ρ的关系

基于这些假设，我们可以得到请求必须在队列中等待的平均时间的一个简单结果：

$$W = \rho \times b/(1-\rho) \tag{7.4}$$

这里，ρ指λ与μ的比率。

参数ρ是服务器的**利用率**（utilization coefficient）。对于任何时间段，这一比率等于服务器的繁忙时间和整个时间段的长度之比。

图7-5显示了平均等候时间W与ρ的关系。可以清楚地看到，参数ρ在排队过程中起了重要的作用。如果ρ接近于零，平均等候时间也接近于零。这意味着请求不必在缓存中等待（它在请求到达时是空闲的），请求可以一也确实立即进入服务器。但是，如果ρ趋近于1，那么，等候时间快速增加。这一关系具有非线性特性。队列的这种行为从感觉上也非常明显，因为ρ代表了输入流的平均速率和服务的平均速率的比。分组间间隔的平均值与平均服务时间越接近，服务器处理负载就越困难。

7.3.2 作为分组处理模型的M/M/1

图7-6描述了之前介绍的模型元素和分组交换网络元素之间的关系。

- 到达交换机输入接口的分组流对应于服务请求的输入流。参数λ对应于分组到达率。
- 交换机输入接口的缓存对应于M/M/1模型的缓存。
- 处理分组、并将它们发送到输出接口的处理器对应于服务器。从输入缓存到输出信道的平均分组转发时间对应于平均请求的服务时间。参数m对应于资源（交换机的处理器或接口）的性能。

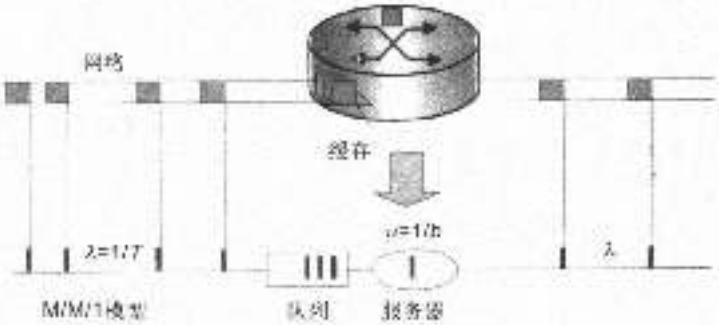


图7-6 M/M/1模型和网络元素的对应关系

我们需要指出的是，之前描述的模型是网络中实际发生的过程的一种简化描述。分组处理的许多典型的特点（交换机缓存的有限容量、将分组储存的入缓存的时间不为零，等等）并没有在这一模型中加以考虑。但是，它的价值在于它揭示了排队行为的量化特性。因此，对于理解影响队列长度的主要因素，它是比较有用的。

网络工程师非常熟悉图7-5所示的图表。他们将这一图表解读成网络延迟与网络负载的关系。之前描述的模型中的参数ρ对应于参与流量传输的网络资源的利用率。例子包括交换机的接口、交换机的处理器、信道和共享介质。

图7-5还显示出了一些意想不到的东西。很难想像，当服务器或网络资源的利用率接近于1时，他们几乎停止处理负载。毕竟，在这种情况下，负载并没有超出这些资源的能力。相反，它只是趋向于极限。另外，很难解释在ρ的值约等于0.5时，队列还存在。流量的传输能力是负载两倍，但是队列还存在，并且，平均而言，队列包括多个等待的请求（分组）。

这些看似矛盾的结果是随机过程发生的系统的特性。因为λ和μ是较长时间内的平均速率，没有什么可以防止流在短时间内偏离这些值。当分组到达率显著地超出服务率时，队列便产生了。

这一模型带给我们的主要结论是：为了保障高QoS，需要防止网络资源的利用率上升到0.9以上。

资源过载可能造成网络的完全瘫痪。如果这一情况发生，有效数据传输率可能为零，虽然网络还在继续传输分组。如果所有分组的传输延迟超过一个特定的阈值，这一情况就会发生。正因为这样，目的节点丢弃了所有这样的分组，因为他们超出了时限。如果网络中运行的协议使用基于确认和分组重传的可靠数据传输过程，这一拥塞过程将像雪崩似地发展。

另一个重要的参数直接影响了分组交换网络中的排队过程，它就是在输入分组流（即，突发的到达流量）中，分组间间隔的变动。我们基于输入流服从普阿松分布的假设，分析了M/M/1模型的行为。这一分布具有相当大的延迟标准差（我们提到过，假如平均间隔等于T，变动率为1，那么，平均变动等于T）。如果输入流的分组间间隔变动减小的话，会发生什么情况呢？或者，如果输入流突发性特别强（即标准差等于10，甚至是100）的话，会有什么效果呢？

不幸的是，对于这种情况，排队理论的模型不能提供类似于公式7.4那样简单的相关性。因此，为了获得结果，用户必须使用网络模拟的方法，或者在真实的网络中进行实验。

图7-7显示了模拟模型产生的一组曲线，它们对应于W和ρ的关系，根据输入流变动率（CV）的不同值获得。这一模拟模型考虑了真实网络中存在的固定延迟。CV参数等于1的那条曲线对应于普阿松输入流。从图示中可以看到，输入流的突发性越小（CV趋向于0），当利用率接近于1时，雪崩式的生成队列的效果越小。另一方面，CV值越大，这一过程出现的时间越早（并且ρ的值越小）。

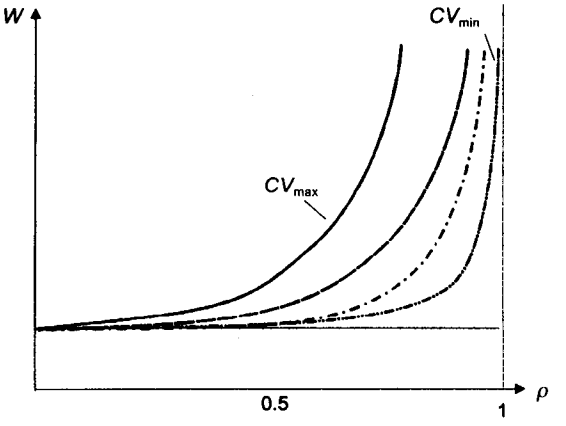


图7-7　流量突发性对延迟的影响

通过分析图7-7所示图表的行为，可以得到两个结论：

- 仅仅是关于资源利用率的信息（ρ）不足以确定网络交换机中队列的延迟。为了得到一个更准确的估计，我们需要知道流量突发性参数。
- 为了改善QoS（减小延迟程度），我们需要设法平缓流量（即，减小流量的突发性）。

7.4 QoS机制

7.4.1 在低负载方式下运行

通常，网络上同时传输许多信息流。每一个流根据特定的QoS要求请求服务。每个流在从源节点到目的节点的路由上，通过多个网络交换机。在每一个交换机上，它经过两个队列——到交换机处理器的队列、到交换机输出接口的队列。你已经知道了直接影响延迟值，也就是影响网络QoS的最重要因素是资源利用率。因此，为了保证所需要的QoS，我们需要保证，在流的路由上，服务流的每一个资源的利用率都不超过预定义的值。

保证满足所有流QoS要求的最简单方法是将网络运行在低负载模式，这时，交换机的所有处理器和接口只使用最大性能的20%到30%。

但是，这抵消了分组交换网络最大的优势，也就是，它在传输突发流量时的高性能。

7.4.2 不同的服务类别

在**重负载网络**（heavily loaded network）中支持QoS是一个困难但重要的任务。在这种情况下，网络中存在不同类别的流对我们是有帮助的。为了简化起见，我们将所有的流分为两类：

- **延迟敏感的流量**（delay-sensitive traffic）（实时流量或同步流量）
- **弹性流量**（elastic traffic）可以容忍较大的延迟，但仍然对数据丢失敏感（异步流量）

我们不知道延迟和资源利用率的确切相关关系，但是，我们知道它们的一般相关性。如果对于延迟敏感的流量，我们保证每一资源的利用率不大于0.2，很显然，每一个队列中的延迟将非常小。这些延迟应该可以被大多数此类应用所接受。弹性流量可以允许一个更高的利用率（虽然它还是不能超过0.9）。为了确保这一类的分组不丢失，需要提供足够的缓存容量，用来存储突发期间所有到达的分组。图7-8显示了这种负载分布的效果。

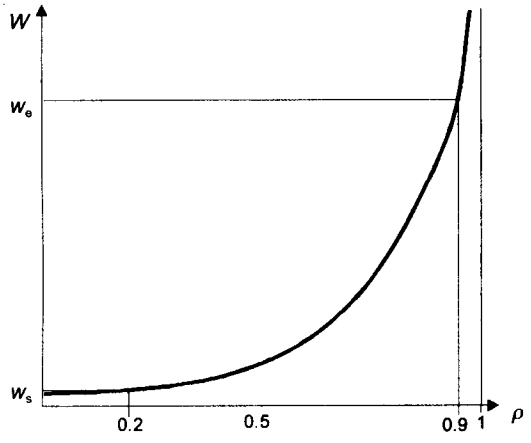


图7-8 服务弹性流量和实时流量

延迟敏感流量的延迟等于w_s；弹性流量的延迟等于w_e。

很长时间以来，分组交换网络只传输弹性流量。因此，主要的QoS要求包括最小化分组丢失、提供一些方法保障每一网络设备的利用率不超过0.9。解决这一任务的方法称为拥塞控制方法。

从20世纪90年代早期开始，传输延迟敏感数据开始变得必不可少，于是，情形就更复杂了，我们需要寻找新的方法。也正是在这段时间里，术语服务质量出现了。它反映了不同类型流量更详尽、更差异化的要求。

为了实现这两类流量不同的资源利用率值，我们需要在每个交换机中支持两种队列。从队列中提取分组的算法必须给予延迟敏感分组队列一定的优先权。如果这一队列中的所有分组都受到优先的服务，另一个队列中的分组只有在第一个队列空时才受到服务，那么，第二个队列就不会影响到第一个队列，对第一个队列而言，仿佛它是不存在的。因此，如果优先流量的平均速率（λ_1）和资源性能（μ）的比等于0.2，那么，这一流量的利用率也等于0.2。

对于弹性流量，它的分组处理总是需要等到处理完所有的优先分组，所以，利用率必须以另一种方式计算。如果弹性流量的平均速率等于（λ_2），那么这一流量的利用率等于（$\lambda_1+\lambda_2$）/μ。因此，如果我们需要保证弹性流量的利用率等于0.9，它的强度必须根据以下计算获得：$\lambda_2/\mu=0.7$。

作为所有QoS支持方法基础的一般思想是：每一资源的总性能必须非均匀地分配给不同类型的流量。

我们可以引入两种以上的服务类型，并设法保证每种类型都可以根据它的利用率受到服务。这一任务被解决后，对QoS特性的进一步改进可以通过使用其他的方法（例如，通过降低流量突发性）来实现。

现在，我们需要找出，在每一个网络节点中，如何保障不同类型的流量。

网络开发者在整个分组交换网络的发展过程中都在设法解决这一问题。实现这一目标可以使用多种机制的不同组合。

7.5 队列管理算法

当拥塞发生时，网络设备不能以分组到达的速率，向输出接口传输分组，这时，我们需要进行队列管理。如果这一过载是由网络设备处理单元的性能不足所造成的，那么，相应输入接口的输入队列用来临时存储未处理的分组。当我们将请求划分为不同类别时，同一个接口中可能有多个输入队列。如果过载由输出接口的带宽不足所造成，那么分组暂时存放在输出接口的队列中。

7.5.1 FIFO算法

传统的**先进先出算法**（FIFO algorithm）的本质是，如果发生过载，所有的分组被放进一个公共的队列，然后根据它们到达的顺存从队列中提取——也就是，最先进入队列的分组最先出队。在所有的分组交换设备中，FIFO算法都是默认的使用方法。它的优点包括易于实现、不需要配置。但是，它也有相当严重的不足——不能对属于不同流的分组区别处理。所有的分组被放入公共队列，具有相同的优先级。这包括延迟敏感的语音流量分组和备份数据分组，后者对延迟不敏感，但是流量相当剧烈，它们较长时间的突发性可能对语音流量造成很长时间的延迟。

7.5.2 优先权排队

优先权排队算法（priority queuing algorithm）在计算的许多领域都很流行——例如，在多任务操作系统中，某些应用必须比其他的应用有更高的优先权。这一算法也用于优先权排队，其中，某些类型的流量必须比其他的流量有更高的优先权。

优先权排队机制基于将所有网络流量分成一些类，对每个类赋予某个数值特性，称为**优先权**（priority）。

流量分类（traffic classification）是一个单独的任务。分组可以根据不同的特性归类于不同的优先权类别：目的地址、源地址、产生这些流量的应用的标识符、或是这些分组头部包含的其他特性的组合。分组分类的规则是**网络管理策略**（network management policy）的一部分。

流量分类点（traffic classification point）可以位于任何通信设备中。扩展性更好的解决方案将流量分类的功能赋予一个或多个专门的设备，这些设备位于网络的边缘。例如，这一功能可以被赋予企业网交换机，这些交换机连接着终端用户的工作站，或是连接着服务提供商网络的边缘路由器。在这种情况下，分组中需要包含一个特殊的域，它用来存储分配到的优先权值，这样，对设备分类后，所有对流量进行处理的网络设备都可以使用这一信息。大多数协议的分组头部提供这样的域。当分组头部没有特殊的优先权域时，我们需要开发一个额外的协议，它将包含一个新的、提供这样一个域的头部。以太网协议使用了这样的解决方案。

不但交换机或路由器可以指定优先权，运行在源节点上的应用也可以指定优先权。同样需要知道的是，如果网络上没有集中式的策略指定优先权，那么，网络设备可能不同意其他网络节点赋予分组的优先权。在这种情况下，设备将根据本地的策略，重写优先权值。

与所选择的流量分类方法无关，支持优先权排队的网络设备中有多个队列[1]。这些队列的数量对应于优先权类别的数量。在拥塞时到达的分组根据它们的优先权类别放入队列。图7-9给出了一个使用高、中、正常、低四种优先权队列的例子。如果具有更高优先权的队列还包含分组，那么，设备将不会处理具有较低优先权的队列。因此，具有中优先权的分组总是在高优先权分组队列为空时才得到处理。相应的，低优先权分组只有在所有更高优先权（高、中、正常）的队列为空时才得到处理。

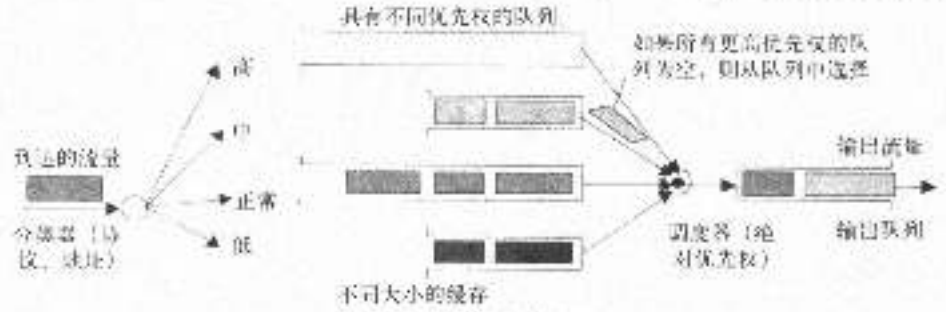


图7-9 优先权排队

通常，所有优先权队列将默认地获得同样大小的缓存，但是，许多设备允许管理员对每一队列单独配置缓存大小。缓存大小决定了可以存储在某一优先权队列中的最大分组数。如果分组到达的时候缓存已满，那么分组就会被丢弃。

通常，缓存大小（*buffer size*）被确定为可以安全地处理的队列的平均长度。但是，估计这个值是相当困难，因为它取决于网络负载。正因为这样，为了实现这一目标，需要在很长一段时间内连续检测网络运行。一般地，流量对用户的重要性越高，它的速率和突发性越高，那么，它所需要的缓存大小就大。在图7-9所示的例子中，高优先权和正常优先权的流量被赋予了较大的缓存，对于另外两种类别，分配了较小的缓存。对于高优先权流量，这种解决方案的目的非常明显，对于正常优先权的流量，我们预计它将具有较高的突发性，流量较宽。

优先权排队方法保证了最高优先权队列中分组的高QoS。如果这些分组到达设备的平均速度不超过输出接口的带宽（以及设备的处理器性能），那么具有最高优先权的分组总是得到它们所要求的带宽。这些分组的延迟程度也达到最小。但是，它并不为零。主要取决于这些分组流的特性。流的突发性和信息率越高，这些高优先权分组形成队列的可能性越大。其他所有优先权的流量对于高优先权分组几乎是透明的。请注意我们说的是“几乎”，因为可能存在这样的情况，具有较高优先权的分组必须等待设备完成对低优先权分组的处理。当高优先权分组到达时，设备正好开始处理低优先权分组，这种情况就会发生。

对于其他优先权类别，提供给它们的QoS将比最高优先权分组低。请注意，这一质量的降低程度很难预测。如果最高优先权的流量相当强烈，这一降低将会非常明显。如果在某些时段，仅由最高优先权流量决定的设备利用率上升到接近1，那么，在这些时间段中，所有较低优先权类别的流量将接近于停滞。

正因为这样，优先权排队适用于网络中存在某一类实时流量，但是它的强度并不高的情况。因此，服务这一类的流并不会损害对其他流量的服务。例如，语音流量对延迟相当敏感，但是它的速率很少超过8~64Kb/s。因此，如果这一流量被赋予最高优先权，提供给其他类别流量的服务并不会受到显著的影响。但是，也可能存在其他的情形。例如，一个网络需要传输视频流量，它需要一个高服务优先权，但是速率相当高。对于这种情况，我们开发了特殊的排队算法，保证

[1] 有时候，多个队列由一个队列代表，它包含了不同类别的请求。如果请求根据它们的优先权从队列中登录，那么它只是同一个机制的另一种表示方法。

即使高优先权流量的速率显著上升，低优先权的流量也会受到服务。下一小节将介绍这一算法。

仔细的读者可能已经注意到，在介绍优先权排队时，我们是基于流量的类别，而不是单独的流。这一特性相当重要，它不仅与优先权排队算法有关，还与其他保障QoS的机制相关。

网络可以以不同的粒度（granularity）对流量提供服务。一个单个的流是QoS机制考虑的最小服务单位。

如果我们对每个流保障单独的QoS参数，那么，我们就是在流的层次上处理QoS。

如果我们在保障QoS参数时，将一个公共聚集流中的多个流结合起来，不再区分单个的流，那么，我们就是在流量类别的层次上处理QoS。这些类别也称做**流量聚集**（traffic aggregate）。

要点　为了将多个流结合成一个聚集流，需要保证它们有相同的QoS要求，进入网络和离开网络时，有共同的输入点和输出点。

7.5.3　加权排队

加权排队算法（weighted queuing algorithm）是为了向所有类型的流量提供某一确定的最小带宽、或至少遵守某些延迟要求而开发的。类型的权重是保留给这类流量的资源（例如，处理器或交换机的输出接口）总带宽的某一百分比。

与优先权排队一样，加权排队需要将数据分成多个类。对于每一个类，创建一个单独的分组队列。但是，在加权排队中，每一个队列被赋予某个资源带宽的百分比，而不是优先权。在资源过载时，会保障提供这一百分比的带宽给这类流量。对于输入流，资源的角色由处理器担任，对于输出流（完成交换以后），资源的角色由输出接口担任。

示例　图7-10显示了一个网络设备支持5个到输出接口队列的例子。在拥塞的情况下，这些队列分别被赋予了输出接口总带宽的10%、10%、30%、20%、30%。这一目标通过循环服务队列的方式实现。在每一个服务周期，从每个队列中提取特定数量的分组，对应于队列的权重。如果在这个例子中，队列查找周期是1秒钟，输出接口的速率是100Mb/s，那么，在拥塞的情况下，第一个队列将获得10%的时间（即，100msec），10Mb的数据将从这一队列中提取。从第二个队列中提取的数据量也是10Mb，相应地，从第三个队列中获取30Mb的数据，从第四个队列中获取20Mb的数据，从第五个队列中获取30Mb的数据。

因此，对一类流量都将得到要保障的最小带宽。在大多数情况下，这一结果比用高优先权类别压制低优先权流量更容易被人接受。

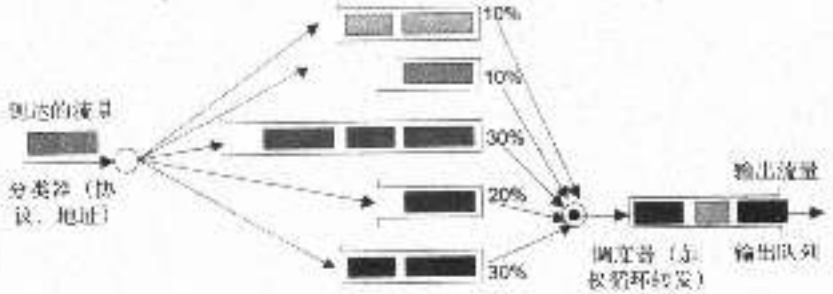


图7-10　加权排队

由于数据以分组的形式从队列中提取，而不是以比特的形式，所以，流量类之间的实际带宽分布总是与计划有些差别。例如，在网络拥塞的情况下，第一类流量可能是9%或12%，而不是10%。

每个循环周期的时间越长，流量类间满足所要求的比例就越准确。这是因为从每个队列中选择了大量的分组，因此，分组大小的影响就减小了。

另一方面，较长的循环周期会造成较大的分组传输延迟。例如，在上面的例子中，周期等于1秒，延迟可能大于1秒，因为仲裁器以不超过每秒1次的频率访问队列。此外，每一队列可能包含多个分组。因此，在选择循环周期的长度时，我们需要保证带宽比例的精确性和减小延迟之间的平衡。

对于上面的例子，1 000μs的循环周期可以保证这样的平衡。一方面，它确保了每一类的队列都可以在每1 000μs中得到一次服务。另一方面，这一时间足够从每一个队列中提取多个分组。在我们的例子中，第一个队列将有100μs的时间，足够向网络传输一个快速以太网帧或10个千兆以太网帧。

在使用加权排队算法的时候，特定类型流量的资源利用率极大地影响了这一类中分组的延迟和延迟变动。在这种情况下，利用率是某类流量的输入速率和按权重分配给它的带宽之比。例如，如果我们将输出接口总带宽的10%分配给第一个队列（即，10Mb/s），进入这一队列的流的平均速率为3Mb/s，那么，这个流的利用率是3/10=0.3。图7-5所示的相关性显示出，对于这一利用率的值，延迟将会很小。如果这一队列的输入流为9Mb/s，队列就会显著增长。如果超出了10Mb/s的限制，由于队列溢出，部分分组将被丢弃。

队列和延迟的定性行为看上去和FIFO队列相似——就是说，利用率越小，平均队列长度和延迟就越小。

和优先权排队一样，在使用加权排队时，管理员可以为不同的队列类型手工分配不同的缓存大小。在拥塞的状况下，减小队列的缓存将造成更多的分组丢失。但是，没有丢弃的分组的等候时间将缩减。

另外，还存在另一种类型的加权排队——**加权公平排队**（weighted fair queuing，WFQ）。这时，资源的带宽被所有的流平分（即公平性）。

> **注意** 只有发生拥塞，每个队列保持为满的状态时，加权排队才保证不同队列流量速率之间的关系。如果有一个队列是空的（这意味着对于这类的流量，当前并没有发生拥塞），那么，在当前的轮循查找中，这一队列就被忽略了，用来服务这一队列的时间根据其他队列的权重，被分配给其他所有的队列。因此，在某一段时间中，某一类流量的速率可能高于预分配的输出接口带宽的百分比。

7.5.4 混合的排队算法

之前介绍的两种方法各有优点和缺点。优先权队列保障了最小延迟程度，至少对于最高优先权的流量是这样的。这一算法首先服务最高优先权队列中的流量，无论它们的速率有多高，也不保证任何低优先权队列中流量的平均带宽。

加权排队保障了平均流量速率，但不提供任何和延迟有关的保证。

混合排队算法试图在这两种方法之间找到一种折衷。这类算法中最流行的一种使用一个优先权队列，根据加权算法服务其他所有的队列。通常，优先权队列用于实时流量，其他队列用于多种类型的弹性流量。每一类弹性流量在拥塞发生时，获得一些受保障的最小带宽。这一最小值根据服务完优先流量后剩下的带宽的百分比计算得出。显然，我们需要限制优先流量，以防它会消耗掉资源的所有带宽。通常，这通过使用**流量轮廓工具**（traffic profiling tool）来实现，我们将在本章的后面对此进行介绍。

7.6 反馈

7.6.1 目的

队列管理算法是防止网络拥塞的必备工具，但是这些工具并不够。这些算法照目前的样子理解当前的情形，并尽力去改进，以达到资源短缺情况下的公平资源分配。但是，它们不能消除带宽不足。这些机制属于为**拥塞控制机制**（congestion control mechanism），当网络已经工作在拥

常状态下时，这些机制将被启动。

另一类的工具试图预测并**防止网络拥塞**（prevent network congestion）。这些工具被称为拥塞避免机制。这类工具的主要目的是防止出现拥塞的情形，因为以较低的速度无拥传输数据，要比以较高速度传输，但在拥塞期间丢失分组好得多。

只有在每个网络交换机每个接口所传输的所有数据流的总速率小于接口带宽时，防止网络拥塞才是可能的。可以使用两种方法来实现这一目标：或者是增加接口的带宽，或者是降低流的速率。第一种方法与网络设计和规划相关，我们在此不做考虑。

第二种方法可以使用两种不同原理的方法实现。其中之一利用了**反馈机制**（feedback mechanism）。拥塞的网络节点使用这一机制，请求流量路由（或是属于同一流量类别的多个流）上的前一个节点暂时降低流量速率。当这一节点的拥塞结束后，它发送另一个消息，允许增加数据传输速率。这一方法不需要了解之前的流量强度。它只是简单地对拥塞做出反应。这一方法还假设，运行在所有节点上的协议将对通知它们拥塞的消息做出反应，并相应地降低流量速率。

另一种方法是对经过网络的流**预留**（reserving）带宽。这一方法需要预先知道流的速率，如果流的速率发生了变化，还需要知道对这一信息的更新。在下一小节中，我们将更详细地介绍资源预留的原理。目前，让我们集中讨论反馈机制。

7.6.2　反馈参与者

现在有几种反馈机制。它们的不同之处在于反馈提供的信息，产生这一信息的节点类型，对这一信息做出反应的节点类型：终端节点（计算机）或中间节点（交换机或路由器）。

图7-11显示了可以用来组织反馈的不同方法。

反馈1在两个网络终端节点之间组织。这是降低网络负载的最基本的方法，因为只有终端节点（发送端）可以降低向网络发送信息的速率。但是，这一反馈类型不是标实控制方法，因为它的主要目标是降低目标节点的负载，而不是网络设备的负载。由于这一问题的产生是因为分组到达网络资源的速度暂时高于资源可以处理它们的速率，因此，从原理上说，这是同一个问题。但是，在这种情况下，并不是交换机担任了网络资源的角色，而是终端节点。传统上，这种反馈类型被称为流控制。网络设备并不参与这种反馈机制的运行。它们只是在终端节点之间传输适当的报文。尽管名字不同，拥塞控制方法和流控制方法使用了共同的机制。

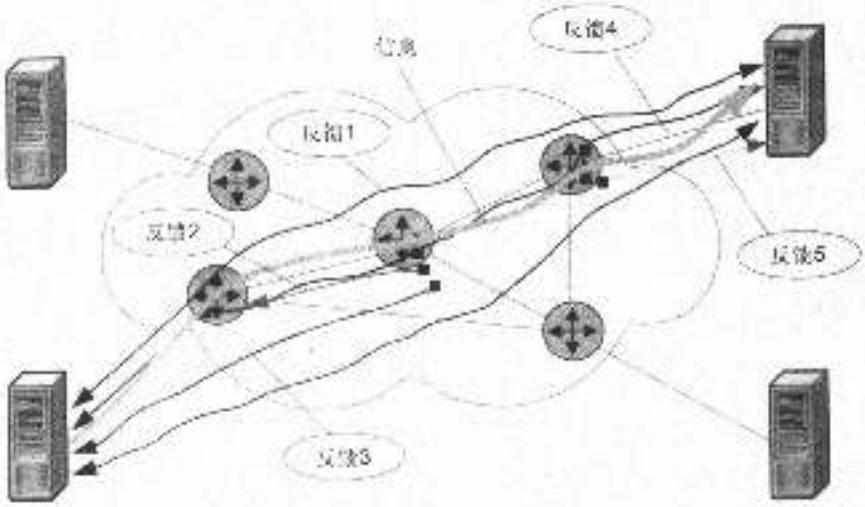


图7-11　反馈参与者

在组织反馈时，我们需要计算在网络上传输信息所需要的时间。高速广域网中，在目标节点传送消息通知源节点目标节点已过载的这段时间中，源节点可能已经传输了数千个分组。因此，并不能及时地消除过载。

根据自动控制理论，和原来的意图相比，反馈回路中的延迟可能造成许多意想不到的结果。例如，系统中可能出现摆动的进程，防止系统到达平衡状态。在因特网发展的早期，这种现象很常见。例如，由于反馈和路由算法的不完善，过载的区域常常出现，并周期性地在网络上移动。造成这一问题的原因很明显：反馈回路中的延迟将过时的被控制元素的状态提供给了控制的元素。

在这种情况下，源节点得到的目标节点队列状态信息具有延迟。因此，可能当源节点开始减小传输速度的时候，由于目标节点上没有队列，所以并不需要任何减小。有时候，当源节点收到延迟的信息，开始增加信息率的时候，目标节点正开始经历过载。为了消除这种效果，反馈回路中通常包括一个集成的元素。这一集成的单元不但考虑当前的反馈消息，也考虑了之前的多个消息，以此来决定情况的动态变化，并做出相应的反应。

在两个相邻的交换机之间组织反馈2。交换机通知它上方（就某一发送流而言）的交换机它正经历拥塞，缓存已经被填充到临界值。接收到这一消息后，上游邻居必须在拥塞交换机的方向上暂时降低数据传输速率，这样，就消除了拥塞问题。这一解决方法对网络整体并不是很有效，因为流会以和原来一样的速率离开源节点。对于经历拥塞的交换机而言，由于它有时间降低溢出的队列，这仍是一个好方法。但是，问题被转移到了上游交换机，由于它开始以更低的速率从它的缓存中传输数据，它可能会出现过载。这一方法的优点是较小的反馈延迟，因为两个节点是邻居节点，当然，前提是它们不能通过卫星信道连接。

在中间交换机和源节点间组织反馈3。虽然反馈消息通过多个中间交换机向源节点方向传输，但这些中间交换机并不对此做出反应。

反馈4是最常见的情况（图7-11）。这时，与第一种情况相似，目标节点生成过载消息，向源节点传送。然而，这一情况的不同之处在于，每一个中间交换机对这一消息做出反应。首先，这一方法降低了向目标节点方向的数据传输速率，其次，这一方法允许在反馈回路上改变消息的内容。例如，如果目标节点要求降低速率到30Mb/s，那么，中间节点在衡量了它自己的缓存状态后，可能要求降低速率到20Mb/s。这一方法的多样性在于，不但目标节点可以生成反馈消息，任何中间交换机也可以生成反馈消息。

在介绍组织反馈的不同方法时，我们假设过载消息的传输方向与用户信息的传输方向相反。然而，有些通信协议并不提供中间节点生成这些消息的能力。在这种情况下，我们使用了一种构造的技术——将关于拥塞的消息传输到目的节点，目的节点将它转换成反馈消息，并向所要求的方向发送（即向源节点方向）。图示中的反馈5显示了这一方法。

7.6.3 反馈信息

目前使用的反馈方法使用了如下类型的信息：

- 拥塞指示
- 最大传输速率
- 最大数据量（信用量）
- 隐含信息

拥塞指示（congestion indication）。它不包含网络或节点的拥塞程度。它只是简单地报告了拥塞。节点收到这一消息后做出的反应可能是不同的。在某些协议中，节点必须在某个方向上停止传输信息，直到它收到另一个反馈消息，允许它继续传输。在其他的协议中，节点具有适应性的行为：它将传输速度降低某个值，等待网络回复。如果带有拥塞指示的反馈消息继续到达，节

点继续降低传输速率。

最大传输速率（maximum transmission rate）。第二种类型的消息指示出源节点或中间节点必须遵守的速度阈值。和前一种方法相比，它是一种更精确的拥塞控制方法，因为它明确地通知上游邻居传输速率必须降低到什么水平。它必须考虑网络的消息传输时间，以消除网络中的摆动，这样，就保障了所要求的对拥塞的回复。因此，在广域网中，这一方法通常作为例子中的反馈4来实现，使用网络中的所有交换机。

最大数据量（maximum data volume）。最后一种消息类型和分组交换网络中广泛使用的滑动窗口算法有关。滑动窗口算法之前已经做过介绍。这一算法不但保障了可靠的数据传输，还提供了组织流控制反馈（如果反馈在终端节点之间组织）或拥塞控制反馈（如果反馈在网络交换机之间组织）的能力。

带有反馈信息的参数是当前窗口大小（在介绍这一机制的工作原理时，我们使用W来指代）。大多数实现滑动窗口算法的协议提供了在确认中指出当前窗口大小的能力，以确认接收到下一段数据。这一当前窗口大小的信息对应于接收节点的当前状态。

这一参数也被称为信用量，由接收节点提供给发送节点。发送节点可以发送特定信息量（或是特定数量的分组，如果窗口大小以分组来衡量），对应于信用量。但是，如果信用量用完了，发送节点就无权发送任何信息，直到它接收到下一个信用量。在拥塞的情况下，接收节点减小窗口大小，这样就减小了负载。如果拥塞被消除，接收节点再次增加窗口的大小。

这一算法的可用性受到一些限制，因为它只能运行在面向连接的协议中。

隐含信息（implicit information）。这一方法中，发送节点根据某些隐含指示，而不是直接的反馈消息，来决定接收节点是否处于过载中。例如，分组丢失可以作为一种隐含指示。为了能检测出分组丢失，协议必须是面向连接的。在这种情况下，超时、或是重复的肯定确认，可以被理解为分组丢失的隐含证据。但是，分组丢失并不总是网络拥塞的证据。事实上，网络拥塞只是分组丢失可能原因之一。其他分组丢失的原因包括通信设备的不可靠运行等，这些包括硬件失效和噪音造成的数据扭曲。然而，由于对于拥塞和不可靠网络运行的反应必须是一样的，即降低传输速率，分组丢失原因的不确定性不会造成任何问题。

使用隐含拥塞信息的协议例子是TCP协议。这一协议在控制流时，使用了明确的反馈信息，在控制拥塞时，使用了隐含信息（分组丢失、重复确认）。对于前一种情况，源节点将窗口大小设置成目标节点所指定的值。对于后一种情况，源节点自行决定需要将窗口减小到什么程度，以减小网络拥塞的效果，或对低传输可靠性做出反应。

7.7　资源预留

7.7.1　资源预留和分组交换

资源预留是反馈机制的替代。这一机制也被归类为拥塞避免工具。然而，资源预留机制试图将拥塞程度限制到某种可接受的值，而不是在拥塞发生时做出反应。这一值必须保证网络交换机中实现的拥塞控制算法能处理短期的过载，提供所需要的QoS参数值而不使用反馈机制。

分组交换网络中的资源预留在原理上不同于电路交换网络中的类似过程。在电路交换网络中，物理信道带宽的固定部分被预留给每一个连接（电路）。流以等于被预留带宽的固定速率在网络上传输，这一速率等于最大流的速率。电路带宽总是为这一个流所预留，不能动态地重新分配给其他的流。如果不事先预留资源，电路交换网络就不能工作，这是它们的基本原理。

在分组交换网络中，资源预留并不是必须的。然而，如果分组交换网络实现了资源预留，这一过程至少在两方面与电路交换网络的资源预留不同：

- 预留是为了平均流速率
- 带宽可以在不同的流之间动态重新分配

预留要求在流路由上的所有资源检查持续流速率，确保它没有超过资源的性能。如果满足了这一条件，那么，每一个资源都将记住它将传输这一流，资源性能的某些部分将被分配给这一流。

示例 让我们来考虑一个资源预留过程的例子（图7-12）。假设在起始状态，网络资源并没有被预留。后来，我们决定分配一些网络资源给流1。为了实现这一目标，我们至少需要知道所要求的流的平均速率。假设流1的持续速率为15Mb/s，所有通信线路（包括交换机接口）带宽的值为100Mb/s。为了简化起见，每一个输入接口配备有一个专用的处理器，它的性能超出了接口的带宽。

假如满足了这些要求，处理器不会成为瓶颈。因此，在做出资源分配的决定时，我们只需要考虑输出接口的带宽。

可以对流1进行服务，因为在它路由上的所有接口具有足够服务这一流的带宽（15<100）。因此，执行了资源预留，流路由上的每一个接口记住它将15Mb/s的带宽分配给流1。

假设过了一段时间，流2需要预留资源，它具有70Mb/s的持续速率。这一预留也可以完成，因为流2路由上的所有接口的可用（没有为另一流预留）带宽大于70Mb/s。流1和流2都经过的接口（接口i3/S2和i1/S3）具有85Mb/s的空闲带宽，其他所有接口有100Mb/s。在预留之后，接口i3/S2和i1/S3各有15Mb/s空闲带宽。

为流3预留带宽也可以成功。这一流的持续速率是10Mb/s。但是，流4具有20Mb/s的持续速率，为它预留资源将失败，因为接口i3/S2和i1/S3各自只有5Mb/s空闲带宽。

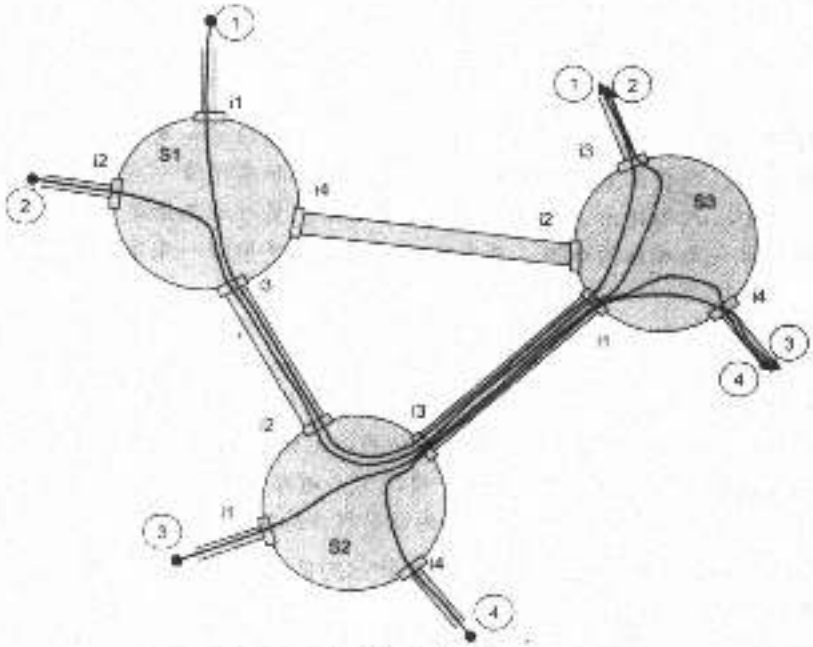


图7-12 分组交换网络中的资源预留

这个例子说明了，如果网络不能保证所要求的QoS水平，网络将拒绝对流的服务。当然，由于我们只集中关注资源预留机制的主要思想，我们简化了这一机制。事实上，网络不但可以保障平

均流速率，而且还可以保障其他QoS特性，例如，最大延迟、最大延迟变动、允许数据丢失的程度。但是，为了实现这一目标，网络必须知道一些额外的流的参数，例如，最大突发程度，以预留所需要的缓存大小。

在完成资源预留时，应该分开考虑延迟敏感流量和弹性流量的空闲带宽。为了保证对延迟敏感流量延迟和延迟变动的可接受程度，最大预留总带宽不能超出每一资源总带宽的50%。正如你可能记得的那样，这时，在优先权排队的条件下，这些流量的延迟将会较小。让我们通过刚才的例子来解释一下。假设我们决定为延迟敏感流量预留30%的资源带宽。那么，如果流1和流3对延迟敏感，这一预留是可能的。另一方面，如果流1和流2对延迟敏感，预留无法实现，因为这两个流的总的平均速率是85Mb/s，大于100Mb/s的30%（30Mb/s）。

如果我们假设延迟敏感流量在优先权队列中服务，那么，为弹性流量预留带宽时，我们需要考虑只有部分带宽，也就是延迟敏感流量剩下的带宽，才能为它预留。例如，如果流1和流3对延迟敏感，我们分配给它们所需要的带宽，那么，只有70Mb/s的未预留带宽可以分配给弹性流量。

如果网络中对某些流的资源预留完成后，会发生什么事呢？从原理上说，在分组处理中，没有什么发生变化。具有资源预留和没有资源预留的网络的唯一区别是，具有资源预留的网络负载合理。这样的网络没有总是工作在拥塞模式下的资源。

在突发时间段中，排队机制继续运作，保障临时的分组缓存。由于我们根据平均速率规划资源负载，在突发时间段中，流的速率可能在很短的时间内超出平均速率。因此，拥塞控制方法依然需要。为了在拥塞时保障所要求的平均流速率，可以用加权队列服务这样的流。

分组交换方法的主要优点被保留了下来。如果某一个流没有完全使用分配给它的带宽，这一带宽可以用来服务另一个流。只为所有流中的一部分预留带宽是一个正常的做法。剩下的那部分流会得到无保留的服务，获得尽力服务。空闲的带宽可能被临时用来动态地服务于这些流，而不违反它们服务于预留资源的流的义务。

电路交换网络不能完成这样的资源重新分配，因为它们没有诸如分组这样的独立的、具有地址的信息单元。

示例　让我们通过车辆交通的例子，来解释分组交换网络和电路交换网络中资源预留的主要区别。假设在某一个小镇，当局决定为救护车保障一些特权。在讨论这一项目的过程中，人们提出了两种互相竞争的想法。第一种想法是，在所有的道路上为救护车提供一条单独的车道，其他所有车都不能使用这一车道，即使路上没有救护车。

第二种方法也建议为救护车提供一条单独车道。但是，如果此刻没有救护车的话，其他车辆有权使用这一车道。当救护车出现时，在特权车道行驶的所有车辆必须立即离开。我们可以很容易地发现，第一种方法对应于电路交换网络中的资源预留，因为在这样的网络中，专用车道只能由救护车使用，无论它们是否需要。第二种方法类似于分组交换网络中的资源预留。在这一情况下，道路的吞吐量得到了更有效的使用。但是，这一方法对救护车比较不利，因为非特权车辆对它们造成了障碍。

让我们从这一类比回到分组交换网络，需要指出的是，为了保障每一个流的服务，之前介绍的预留机制并不够。

我们假设我们确切地知道流的平均速率和突发参数。但是，实际上，这样的信息并不总是可靠的，如果流的速率超出资源预留时考虑的速率，那会发生什么呢？在这方面，另一个问题也没有答案——那就是，如何才能确保在流的路由上自动预留带宽？

为了解决我们提出的这些问题，我们需要一个QoS系统，它将包括除排队算法以外的机制。

7.7.2 基于预留的QoS系统

QoS系统具有分布式特性，因为它的元素必须出现在所有执行分组转发的网络设备中：交换机、路由器、接入服务器。另一方面，对于那些支持QoS的单个网络设备，它们的运作必须协调起来。我们需要保证在分组传输的整个路由上，QoS是相同的。这就是为什么QoS系统必须包含集中式管理元素，使网络管理员可以协调在单个网络设备中配置QoS机制的过程。

基于资源预留的QoS系统包括了多种类型的机制（图7-13）：

- 队列服务机制
- 资源预留协议
- 流量调节机制

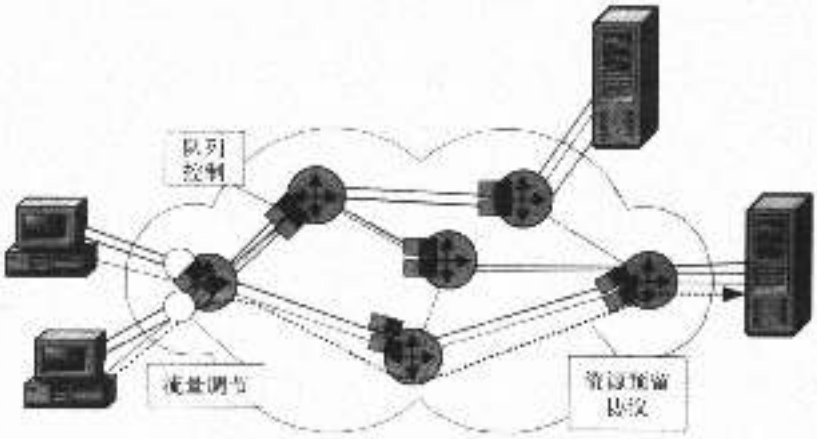


图7-13 基于预留的QoS系统体系结构

队列服务机制（queue-serving mechanism）用于暂时的拥塞时间段。加权排队用于服务弹性流量；对于实时流量，使用了优先权排队。为了减小交换机和路由器上的负载，使用了流量分类服务，因为我们需要支持更少数量的队列，存储更少的流状态信息。

预留协议（reservation protocol）用于在某个流的整个路由上（即基于终端到终端）自动预留资源。预留协议类似于电路交换网络中的连接建立协议。因此，根据这类网络中使用的术语，它们有时被称为信号协议。

资源预留协议在网络上有两次传送。首先，它从信息发送端传送到信息接收端。预留协议的报文表明所谓的**流量轮廓**（traffic profile），它包括诸如平均速率、突发参数、所要求的延迟程度等特性。基于这一轮廓，流路由上的每一交换机决定它是否可以为这一流执行预留。如果它同意执行预留，报文被继续传输，交换机存贮相关信息。如果路由上的所有交换机同意所请求的预留，那么，最后一个交换机发送一个关于资源预留协议的新报文。这一新报文以相反的方向传输。当这一报文传输经过交换机时，每一个交换机登记这一流的预留状态。

不仅仅是终端节点，中间设备也可以发起资源预留协议的运作。在这种情况下，受保障的流服务不会发生在整个流量路由上，而是仅仅在某个网络区域的范围内，这当然会降低提供给流量的QoS。

预留协议既可以为单个流执行预留，也可以为流量类型执行预留。每种情况下的工作原理是一样的。然而，对于流量类型，预留发起者的角色不是由终端节点担任，终端节点主要对它自己的流感兴趣。这一发起者的角色由网络交换机之一来担任。通常，这是服务提供商网络的边缘交换机中的一个，它从不同的用户处接收流。

在具有虚电路的网络中，资源预留协议的功能通常由**建立虚电路**（establishing a virtual circuit）的协议来完成。我们需要指出，连接建立协议本身可能不能完成资源预留，因为这是这一协议的可选功能。在数据报网络中，预留协议是一个独立的协议。这类协议的一个例子是**资源预留协议**（resource reservation protocol, RSVP），它在IP网络中工作。

即使没有预留协议，也可以执行预留。为了做到这一点，网络管理员必须在每个网络交换机中，为每个流手工设备预留参数。

流量调节机制（traffic-conditioning mechanism）关心当前流的参数，确保它们对应于预留时声明的值。它们是一种检查点，在流量进入交换机前进行检查。如果没有这种机制，我们就无法为流量保障所需的QoS，那是因为，如果平均流速率或突发性超出了预留时声明的水平，分组延迟和丢失也将超出流所要求的水平。发生这一情况可能有多种原因。首先，很难对流量参数做出精确的估计。对平均速率和突发性的初步测量可能会产生不准确的结果，因为这些特性随着时间而改变，所以，一周以后，测量的结果可能与现实不符。此外，也不能忽略对流量参数的故意改变，尤其是对于商业服务而言。

流量调节机制通常包括多种功能：

- **流量分类**（traffic classification）。这一功能从到达设备的公共分组序列中选择同一流的、具有相同QoS要求的分组。在具有虚电路的网络中，不需要额外的分类，因为虚电路标识表明了特定的流。数据报网络中通常没有这样的标识符，因此，分类将根据几种分组的特性而被执行：源地址和目的地址、应用标识符等。如果没有分组分类，数据报网络就无法支持QoS。
- **流量监管**（traffic policing）。对于每一个输入流，每一个交换机中都有一套QoS参数，称为流量轮廓（traffic profiles）。监管检查每一输入流和它的轮廓参数之间的一致性。有一些算法可以以分组到达交换机输入接口同样的速度自动完成检查。监管算法的例子包括漏桶和令牌桶算法。我们介绍各种技术（例如，IP、帧中继、ATM）时，将详细地介绍这些算法。

如果违反了轮廓参数（例如，超出了突发尺寸或平均速率），那么这一个流的分组将被丢弃或标记。丢弃部分分组降低了流的速率，使它的参数符合流量轮廓中指定的值。标记分组，或不是丢弃它们，用来保证所有的分组依然被当前节点（或是它的下游邻居）服务——虽然在这种情况下，QoS参数被降低，与轮廓中指定的值不一致（例如，延迟增加）。

- **流量整形**（traffic shaping）。这一功能用来对遭受监管的流量进行时间整形。这一功能主要用来平缓突发的流量，使输出设备的流比输入设备的流有更好的一致性。平缓突发帮助降低了当前设备之后处理流量的网络设备中的队列。它也被用来恢复与流相关的那些应用流量的时间关系，例如语音应用。

流量调节机制可能受每个网络节点的支持，也可能只在边缘网络设备中实现。服务提供商在调节客户的流量时，往往使用后一种情况。

在上面描述的系统中，一方面，保障了平均流速率，另一方面，保障了延迟和延迟变动所要求的阈值，这两个行为之间，存在着重大区别。

在使用加权排队时，通过分配带宽的百分比来保障所要求的平均服务速率。因此，网络可以执行对任何平均流速率的请求，只要它不超过在某个流路由上的网络可用带宽。

然而，网络无法用这样的方法配置优先权队列，以保障它严格满足某些预定义的、对延迟和延迟变动的阈值要求。将分组放入优先权队列只能保证延迟足够小，至少比根据加权排队算法所服务的分组要小得多。但是，很难定量地衡量延迟。那么，服务提供商如何遵守SLA呢？

通常，这一问题通过测量网络流量来解决。服务提供商必须为流量组织优先权排队，使用一个或多个优先权队列，测量真实流量的延迟，使用统计方法处理结果。为了实现这一目标，服务提供商必须为不同的流路由建立延迟分布直方图，对每一类延迟敏感流量确定平均延迟、延迟变

动，最大延迟变动，基于这些特性，服务提供商选择它可以向客户保证的QoS特性阈值。通常，这些值的选择有些保留，这样，即使有一些新的客户，网络也可以遵守所声明的保证。

7.8 流量工程

在考虑基于预留的QoS系统时，我们没有考虑网络上传输的流的路由。更确切地说，我们认为这些路由是预定义的，是一个不考虑QoS要求而做的选择。在预定义路由的情况下，我们设法保证这一路由上的一系列流满足QoS要求。

如果我们假设流量路由不是固定的，而是可以选择的，那么，支持QoS的任务可以更有效地解决。这将允许网络服务更多的流，并保障QoS要求，只要网络的特性（即链路带宽和交换机性能）保持不变。

流量工程（traffic engineering/TE）方法解决了为流（或流聚集）选择路由，满足QoS要求的问题。这些方法同时试图达到另一个目标，平衡所有网络设备的负载，使其尽可能接近最大值。当这一目标实现时，网络将拥有最大的性能，并保证特定的QoS特性。

TE方法也基于资源预留，除了为流选择最优路由外，它们在流的路由上预留网络资源带宽。

对于分组交换网络而言，TE方法相对较新。这主要是因为传输弹性流量没有对QoS参数提出很高的要求。此外，很长时间以来，因特网是一个非商业的网络。因此，最大限度地使用资源并不是作为医特网基础的IP技术的最重要问题。

但现在形势变化了，分组交换网络必须传输多种类型的流量、提供特定的QoS、保障资源的最大利用。然而，为了实现这一目标，我们需要改变一些传统的路由选择的方法。

7.8.1 传统路由方法的不足

基于网络拓扑，而不考虑当前网络负载信息的路由选择是分组交换网络路由协议的主要工作原理。

对于每一源地址—目的地址对，这些协议选择唯一的路由，而对网络上传输的信息流不加考虑。因此，这一对终端节点之间所有的流都在这一路由上传输，根据共衡量标准，这一路由是最短的路由。所选择的路由可能非常合理（例如，计算通信链路的预定带宽，或这些链路带来的延迟），也可能不太合理（例如，只计算源节点和目的节点间中间路由器的数量）。

要点 传统的路由方法认为所选择的最优路由是唯一可能的路由，即使还存在其他的路由（虽然它们可能更长一些）。

具有图7-4所示拓扑结构的网络，是说明这一方法低效率的一个经典例子。尽管交换机A和E之间有两条路由——上面一条通过交换机B，下面一条通过交换机C和D——根据传统的路由原理，从交换机A到交换机E之间的所有流量将通过上面的路由转发。之所以这样做只有一个原因——从中间节点的数量上说，下面的路由更长。因此，这一路由被忽略，虽然它可以和上面的路由并行工作。

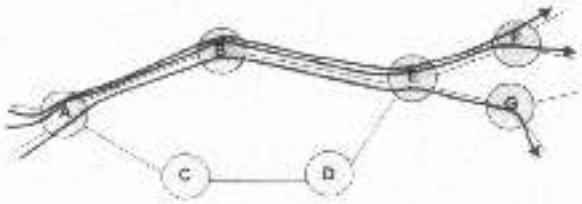


图7-14 选择最短路由方法的低效率

因此，即使最短路由发生了阻塞，分组也会在这一路由上发送。例如，在图7-14所示的网络中，即使上面路由的资源不够服务从A到E的流量，它依然会被使用。同时，下面的路由将被忽略，尽管交换机B和C的资源可能足以高质量地传输这一流量。

网络资源分配方法的低效率非常明显，有些资源工作已过载状态下，有些资源没有被利用。

拥塞控制机制不能解决这一问题，因为当无法找到合理路由时，它们在已经失败的情况下处理这一问题。预留方法也不能解决这一问题，它们不计算网络交换机上的当前负载，试图在选定路由上预留资源。因此，我们需要原理上不同的机制。

7.8.2 流量工程的思想

TE方法使用以下的初始数据：

- 网络的特性——它的拓扑，以及它的交换机性能和通信链路的带宽（图7-15）。
- 负载上的信息——也就是，网络必须在边缘交换机间传输的流（图7-16）。

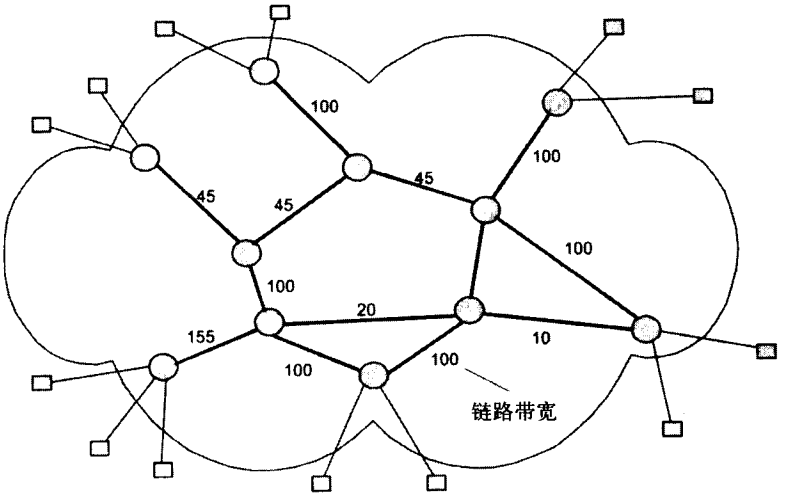


图7-15　网络拓扑和它的资源性能

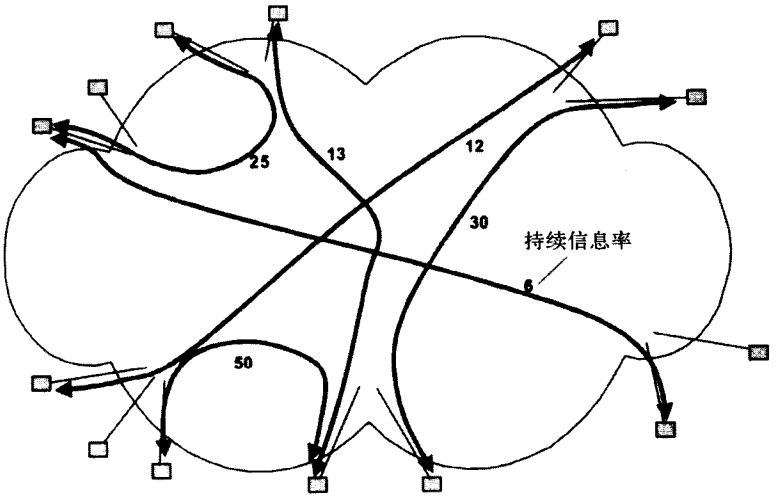


图7-16　负载

假设每个交换机的处理器性能足够服务所有输入接口上的流量，即使流量以最大可能的速率（这一速率等于接口的带宽）到达接口。因此，在执行资源预留时，链路连接起来的交换机的带宽决定了两个接口间的带宽，将这一带宽作为资源。

每一个流由网络进入点、网络离开点和流量轮廓来刻画。为了得到最优的解决方案，我们可以使用对每一个流的详细描述。例如，考虑可能的流量突发性的值、或是分组间间隔的变动。然

面，很难定量地计算它们对网络的影响，而且，这些参数对QoS特性的影响不太重要。所以，为了找到次优流量路由分布，我们通常只考虑流的持续信息率，如图7-17所示。

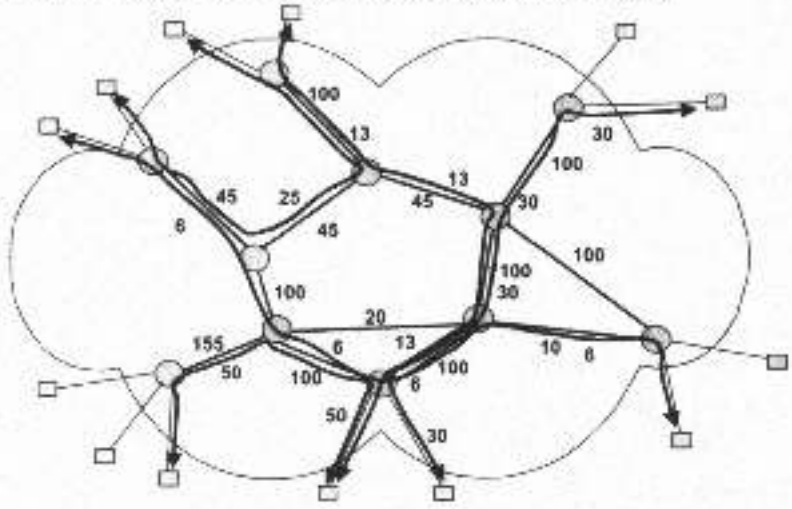


图7-17 网络中的路由分配——针对不同的流来选择路由

TE方法通常用于含有多个流的聚集流，而不是单个的流。由于我们为多个流寻找一条公共的路由，只有那些具有共同的网络进入点和离开点的流可以被聚集。流聚集简化了选择路由的问题，在考虑单个用户流时，由于单个流可能数量相当多，中间交换机必须存储大量的信息。但是，需要指出的是，只有当每一个流有相同的QoS要求时，才能将多个流聚集起来。为了简化起见，我们将在这一小节中使用术语流（flow）同时指代单个流和聚集流，也为这并不改变TE的原理。

TE的目标是决定在网络上传输的流的路由，这意味着，对于每一个流，我们需要找到确切的中间交换机序列来转发，选择路由的方法，必须使所有网络资源达到最大的利用率，并保证每一个流得到所要求的QoS。

最大资源利用率以某种方式来选择，使得拥塞控制机制可以保障所需要的QoS。为了简化起见，从现在开始，我们将假设网络只传输一种类型的流量（例如，延迟敏感流量），然后，在网络必须传输不同类型的流量时，介绍如何推广TE方法。

TE问题的解决方案是对一系列特定流的一套路由，对这些流而言，路由上的所有资源利用率不超过预定义的阈值K_{max}。

图7-17显示了对图7-15和图7-16所示问题的一种可能的解决方案。所选择的路由保证了对于任何流，任何资源的最大利用率不超过0.6（即，在这一例子中，K_{max}等于0.6）。

解决TE问题有不同的方法。首先，我们可以事先在后台模式下找出一个解决方案。出于这一目的，我们需要知道初始数据：网络拓扑和性能、流的网络进入点和离开点、它们的平均速率。之后，在固定的进入点、离开点、预定义的最大利用率条件下，由某些程序来解决如何在数据流间合理分配路由的问题。例如，这样的程序可能通过测试所有可能的情况，找到解决方案。它运行的结果是每个流的确切路由，并标明所有的中间交换机。

第二种方法假设TE问题由网络交换机完成。出于这一目的，我们可以使用标准路由协议的改进版本。路由协议中的改进意味着，交换机和路由器不仅仅互相告知拓扑信息，还交换每一资源上的当前可用带宽值。

在找到解决方案后，我们需要用路由表的形式实现它。在这一阶段，如果你需要在数据报网

络中建立这些路由，可能会出现另一个问题。这类网络中的路由表只处理分组的目的地址。这类网络（例如，IP网络）中的交换机和路由器并不操作流，因为单个流并不以显式的方式存在。在这样的网络中，每个分组在转发过程中是一个单独的交换单元。换句话说，这类网络的转发表只反映出网络的拓扑（对某个目的地址的转发方向）。

因此，将资源预留方法融入数据报网络有不少困难。类似于之前提到的RSVP，资源预留协议除了使用目的地址外，还使用了一系列特性来定义数据报路由器的流。在这种情况下，流的概念只在资源预留阶段使用，在转发的过程中，传统的机制继续工作（即，只使用目的地址）。

现在，我们假设在两个终端节点间有多个流（依然假设所有的流属于唯一的流量类型），并且需要在不同的路由上转发它们。出于使网络负载平衡的目的，而解决TE问题时，这一需求就可能出现。数据报交换机或路由器不能实现这一解决方案，因为对于所有这些流，它在转发表中只有一项记录，对应于属于这些流的分组的公共目的地址。改变数据报网络中交换机和路由器的工作逻辑是不合适的，因为需要根本性的改变。

因此，TE方法目前应用于具有虚电路的网络，对它们而言，实现这一解决方案不会出现任何问题。每一个流（或是同一路由的一组流）接收所分配的虚电路。这些虚电路根据所选择的路由而创建。TE方法成功地运用在ATM和帧中继网络中，这些网络基于虚电路技术。当IP网络运用在ATM或帧中继网络（它们作为基于IP协议的互联网络的一部分）中时，IP网络也可以从TE的优点中获益。一种称为MPLS的新技术专门为了在IP网络中集成虚电路技术而开发。MPLS可以用来在IP网络中解决TE问题。

本书的后一部分将详细介绍每种技术，以及它们的TE方法。

7.8.3　不同流量类别的流量工程

上一小节中考虑的TE问题是一种简化了的形式。所有数据流属于相同的流量类型，具有相同的QoS要求。这允许对所有的流使用相同的最大利用率K_{max}。

可能对每个网络用户而言，存在多种流量类型。我们在考虑资源预留时解释过类似的问题。

这意味着需要对每种流量类型提供一个K_{max}值。

考虑了具有不同QoS要求的TE方法用与单个节点资源预留同样的方式解决了这一问题。如果我们有两种类型的流量，我们必须指定两个最大资源利用率。例如，对于弹性流量，最大利用率不能超过0.9，对于延迟敏感流量，这一值不能超过0.5。通常，预留并不对所有的流都执行，因为部分带宽必须是可获得的。出于这一原因，之前提供的最大值通常会分别降为大约0.75和0.25。

为了达到这一结果，每种资源必须有两个可用带宽的计数器：一个用于延迟敏感优先流量，另一个用于弹性流量。在确定使用某种资源决定路由时，对于优先流量的新的流的平均速率必须和优先流量的可用带宽进行比较。如果可用带宽足够，新的流将使用这一接口通过，那么，新流的速率必须从优先流量可用带宽的计数器中减去，也需要从弹性流量可用带宽的计数器中减去，因为优先流量总是在弹性流量之前得到服务。因此，优先流量也会对弹性流量产生一个额外的负载。如果TE问题是为了解决弹性流量，那么，它的速率将与弹性流量可用带宽的计数器相比较。

如果做了一个肯定的决定，这一速率值只需要从弹性流量计数器中减去，因为弹性流量对优先流量是透明的。

对于每一种流量类型，修改的路由协议必须在网络上传播两个可用带宽参数。如果这一问题被推广为传输多种流量类型，那么，根据网络上存在的流量类型的数量，每一资源必须有同样数量的计数器。路由协议必须传播可用带宽向量，它们包含适当数量的元素。

小结

- 狭义上说，服务质量关注通信设备中的队列对流量传输的影响。当前，QoS方法在分组交换网络技术中占据了重要的位置。如果不实现这些方法，当代多介质应用（例如，IP电话、视频和音频广播、互动远程学习）的运行是不可能的。
- 应用流量可以根据它们对QoS特性的要求，分成两大类：时间敏感流量和弹性流量。
- QoS特性反映了分组在队列中花费时间所造成的负面影响，包括降低传输速率、分组损毁和丢失。
- 优先权排队、加权排队、资源预留和反馈机制允许保障延迟敏感流量和弹性流量的QoS。
- 滑动窗口算法保障了可靠分组传输，也是一个有效的反馈工具。
- 基于资源预留的QoS系统的体系结构包括：
  - 排队机制
  - 预留协议，它们允许为流自动预留所需要的资源
  - 流量调节工具，它们执行流量分类、监管、整形
- 流量工程方法由选择传输流的合理路由构成。路由选择最大化了网络资源的利用，并保障了符合QoS要求。

复习题

1. 请给出产生弹性流量的应用的例子。
2. 对于时间敏感流量来说，哪些QoS特性是最重要的？
3. 在分组交换机中使用队列有什么优点和缺点？
4. 哪个参数对队列大小有最大的影响？哪个参数第二重要？
5. 分组交换网络传输什么类型的流量？它们对网络有什么要求？
6. 优先权排队有什么优点和缺点？
7. 加权排队对什么类型的流量最适合？
8. 是否可能将优先权排队和加权排队结合起来？
9. 具有最高优先权的分组是否可能在队列中延迟？
10. 请列出拥塞控制和拥塞避免的方法。
11. 分组交换网络和电路交换网络中的带宽预留有什么区别？
12. 基于预留的QoS系统有哪些组成部分？
13. 流量工程方法解决什么问题？
14. 在流量工程中，什么流量参数是可改变的？

练习题

1. 假设某些数据流属于CBR类型。数据以125字节大小的分组在100Mb/s的链路上传输。流量轮廓有如下参数：突发时间段的PIR是25Mb/s，分组间间隔的最大偏离是10ms，突发时间段是600ms。如果流量遵守它的轮廓，突发的最大量是多少？
2. 对以下五个流来说，哪个平均花最少的时间到达输出接口的队列中？输出接口为100Mb/s，假设这些流接受加权排队服务，分别分配到接口带宽的40%、15%、10%、30%、5%。平均流速率分别是35、2、8、3、4Mb/s。对所有流来说，分组间间隔的变动系数是相等的。
3. 最高优先权队列中的流，为了以下哪一个事件，必须在队列中等待？

   A. 存在较低优先权的队列

B. 它自己的突发

C. 低优先权流量的突发

4. 有三个队列到具有10Mb/s的输出接口，根据加权排队算法服务。第一个队列中有三个分组，分组1为1 500字节，分组2为625字节，分组3为750字节。第二个队列中有分组4（500字节）、分组5（1 500字节）、分组6（1 500字节）。第三个队列中有分组7（100字节）、分组8（275字节）、分组9（1 500字节）、分组10（1 500字节）。在队列中，分组以升序排列（即第一个队列的第一个分组是分组1；第二个队列的第一个分组是分组4；第三个队列的第一个分组是分组7）。

如果算法的工作周期是10msec，队列分别被分配资源带宽的50%、30%、20%。这些分组将以什么顺序出现在2Mb/s接口的输出部分？在每一周期中，算法总是从队列中提取一个分组（如果它不是空的），即使分组大小表明它的传输将超过分配给这一队列的时间。

5. 用两个周期处理队列会花多长时间（参见前一个问题）？两个周期的时间段中，每一个流被服务的速率是多少？

6. 如何修改问题4中算法的周期时间，使流的速率更接近计划的速率？增加还是降低？

7. 在网络入口处，有些流根据3Mb/s的轮廓，正经受管制。这一流在中间网络交换机中被分配了30%的10Mb/s输出接口带宽。以下哪一个陈述是正确的？

A. 应用这些机制中的任何一个效果都是一样的，因此，不必在交换机上实现资源预留。

B. 应用这些机制中的任何一个效果都是一样的，但是交换机中的资源预留是必要的，因为在网络入口处和交换机中，流会和其他的流竞争资源。

C. 应用这些机制中的任何一个效果都是不同的，在网络入口处，流的速率被限制为3Mb/s；在交换机中，3Mb/s的速率被保证，即使在拥塞期间。

8. 是否可能系统中没有队列，但系统的利用率接近于1？

9. 以下列出的哪些机制需要用来在分组交换网络中，保障高质量的语音流量（64Kb/s的流）传输？

A. 在语音流量路由上的所有交换机中，保留64Kb/s的带宽。

B. 在流量路由上的所有交换机中，用优先权队列服务这一流。

C. 在接收网络节点中使用输入分组缓存。

D. 在路由上的所有交换机的输出队列中，平缓流量。

10. 以下说法是否正确？在分组交换网络中的资源预留剥夺了用户在流之间动态重新分配带宽的能力。

11. 为了避免低优先权流量被高优先权流量所压制，需要应用哪些机制？

第二部分　物理层技术

任何计算机（或者无线电通信）网络的物理基础都是传输链路。没有这些链路，交换机间就不能交换分组，那么诸多的计算机不过是一些昂贵的装备而已。

学习计算机网络设计原则之后，读者或许能够想像计算机网络的一个简单场景——计算机和交换机由不计其数的电缆连接起来。尽管如此，如果我们从更加细致的角度考虑计算机网络，就会发现当我们学习OSI模型时，事情远比它们看起来复杂得多。

专用电缆只用于短距离连接网络设备（例如在LAN中）。当建立WAN或者MAN时，这个方法既代价昂贵又没有效率，这是因为，长距离的传输链路导致了昂贵的价格。更糟的是，为了能够安装这些电缆，还必须获得政府许可。因此，使用现有的地区电路交换网络——使用电话网络，或者使用传输网络——不失为解决WAN和MAN交换机连接问题之道。所以，在电路交换网络中，需要创建电路，它的作用与电缆部分相同。也就是说，电路保证了物理上的点对点连接。当然，电路是一个比电缆复杂得多的技术系统，不过，它的复杂性对计算机网络来说是透明的。专门建立的传输网络用来创建通信链接基础设施，以使这些链路在价格：容量比方面更有效。目前，计算机网络的设计者把64Kb/s到10Gb/s的传输链路作为他们的研究对象。

虽然通信链路的物理属性和技术属性间存在很大差异，但是依然可以用统一的特性描述。任何通信链路在传输离散数据时，其最重要的技术参数都是带宽和容量，其中带宽用赫兹（hertz）来衡量，而容量则是比特每秒（b/s）衡量。容量是指在给定的传输链路上比特流所能达到的最大速率。容量依赖于带宽和编码离散信息所用的方法。

无线链路正在变得十分热门。虽然它只是用来确保计算机网络用户的移动性。而且，无线通信在无法安装电缆或者安装费用很高的情况下是很有用的，这一情况在人口稀少的地区或者是已经铺设电缆但年用严重的地域中比较常见。无线通信使用不同频率的电磁波——无线电波、微波、红外波甚至可见光。无线链路典型的高级别噪声和复杂的传播路径需要使用特殊的信号编码和传输方法。

第8章 传输链路

8.1 引言

建立网络时，可以使用不同的物理介质完成各种传输链路：悬挂在电线杆上的电话和电报线，埋藏在地下的同轴电缆和光纤电缆，连接当代办公室间的铜质双绞线，还有可以穿透一切的无线电波。

这一章涵盖了传输链路独立于它们物理属性的一般特性和这些物理属性之间的独立性。例如：带宽，容量，错误率和抗噪声。带宽是传输链路的基础特性，因为它决定了链路可能的最大信息率，称作链路容量。**奈奎斯特公式**（Nyquist formula）表示了理想链路的可靠性；**香农公式**（Shannon formula）考虑了噪声干扰，表示了真实的通信链路。本章的最后描述了当前电缆标准，这些标准构成了导向通信链路的基础。

8.2 分类

8.2.1 传输网、电路和链路

当描述在两个相邻网络结点间传输信息的技术系统时，会遇到四个术语：链路（link）、电路（circuit）、信道（channel）和线路（line）。这些术语经常被当作同义词，并且在稍加注意的情况下并不会带来什么麻烦。同时，它们的用法有许多独特之处。

- **链路**（link）用来指明两个相邻结点间的部分。这意味着链路不包括转换路由器和多路复用设备。
- **信道**（channel）经常用来指明独立于其他信道的链路带宽部分。例如，传输网络的电路可能由30个信道组成，每个信道保证64Kb/s的带宽。
- **电路**（circuit）是网络末端结点间的复合路径。因此，电路由单独的链路和交换机间的内部连接组成。
- **线路**（line）可以视做上述三个术语的同义词。

虽然有必要理解这些术语用法上的不同约定，但由于实际中这些术语使用上的混乱使我们不必严格地判断它们。这一点在考虑传统电话和新技术领域——计算机网络之间的术语区别时尤为重要。集中的各类网络不断融合的过程使得相关术语混淆问题更加突出。这种情况经常发生，因为在这些网络中，许多机制的使用都是十分通用的，但是其在各自的技术领域却保持着两个甚至更多的名称。

除了这一点，还有诸多客观原因导致了网络术语模棱两可的解释。图8-1表示了一条传输线路的两种变化形式类，在第一类（图8-1a所示）中，线路由几十米的电缆构成。在第二类（图8-2b所示）中，传输线路是一个建立在电路交换网络上的电路。这可能是**传输网络**（transmission network）和**电话网络**（telephone network）两者其中之一。

考虑这样一个例子，计算机网络中的两台交换机由传输网络中的传输链路连接。对于计算机网络，这个连接是一个链路，因为它连接两个相邻结点，并且所有的转换设备对于这些结点都是透明的。然而，对传输网络而言，这一连接是电路——或者更精确地说，是信道——因为它可能使用了每个连接传输网络交换机链路的部分带宽。因此，在计算机网络和传输网络领域的专家之间存在共同的误解，依据也是显而易见的。注意，电话网络本身可能建立在传输网络链路的基础上。

创建传输网络用来为计算机和电话网络提供传输服务。在这些情况下，可以说计算机或者电话网络架设在传输网络上。

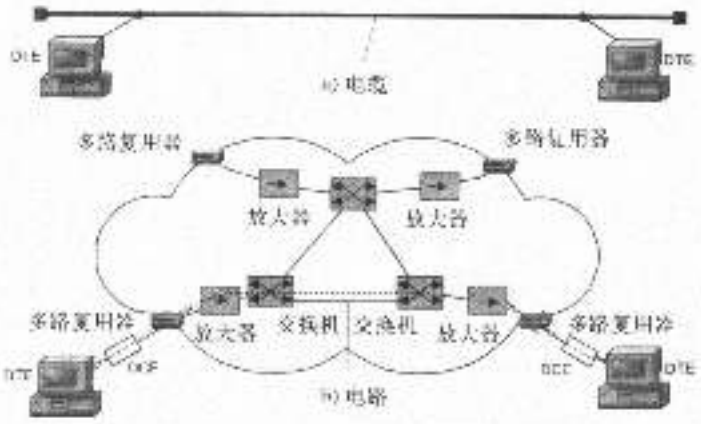


图8-1 通信链路的组成

8.2.2 介质

用于传输信息的传输链路在物理介质上各不相同。

物理介质（physical medium）用于发送传输，可能是一些用于传输信号的电线。明线和电缆通信（地下或者水下）建立在这些电线（图8-2所示）基础之上。信息信号也可以在其他物理介质上传播，例如地球大气层和外太空。第一类使用有线介质（*wired medium*）或导向介质（*guided medium*），第二类为无线介质（*wireless medium*）或无导向介质（*unguided medium*）。

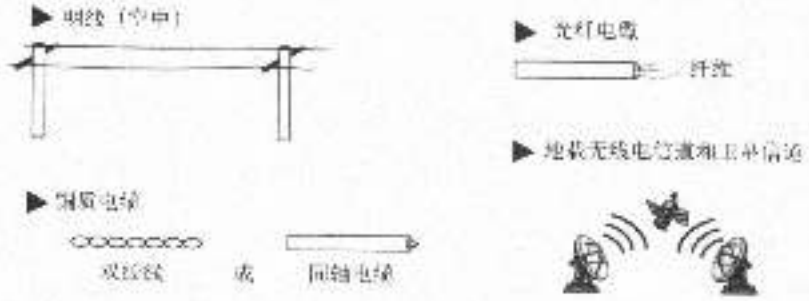


图8-2 传输介质类型

在当代无线电通信系统中，会使用电流或电压、无线电或者光信号传输信息。这些物理过程是不同频率的电磁场振荡。

明线（*Open wire*）（天线）通信线路是指架设在电线杆上没有绝缘层和屏蔽层的电线。目前，这些通信线路仍然用来传输电话和电报信号。现在，这些空中线路迅速被电缆线所替代。尽管如此，如果没有更先进的通信线路可用的话，它们依然可以用来传输计算机数据。但是，这些线路的传输率和抗噪声却不完美。

电缆线（*Cable line*）是一种相当复杂的设计。电缆由包裹多个绝缘（针对电子、电磁、机械、还可能针对气候）层的线组成。电缆线还可以安装在连接不同设备的连接器上。下面三种电缆用在计算机（无线通信）网络中：

- 非屏蔽双绞线（Unshielded twisted pair, UTP）和屏蔽双绞线（shielded twisted pair, STP）

- **同轴电缆**（Coaxial cables）**铜线**
- **光纤电缆**（Fiber-optic cable）

前两种电缆也被称为铜质电缆。

地载和卫星通信无线电信道（Radio channels of ground-based and satellite communication）由无线电波发送端和接收端构成。不同的无线电信道频率和波段不同。广播无线电波段（长波，中波，和短波，也称做调幅或者AM波段）能以非常低的数据传输率确保长距离通信。运行在高频率波段上的信道使用调频（FM），提供稍高的数据传输率。其他频率范围，称做超高频或者微波（300MHz以上），也在数据传输领域使用。频率超过30MHz的信号不会被地球的电离层反射；因此为了保证通信稳定，发送端和接收端必须设置在视线之内。这些频率使用在卫星信道、无线电转播信道、和地区或移动网络，在这些情况下它可以满足条件。

现在，上述提到的所有物理介质类型都在计算机网络中用于数据通信。许多高容量服务由光纤电缆提供，在于它拥有宽广的带宽和优质的抗噪性。因此，光纤电缆用来架构大型区域网络、MAN和高速LAN的主干。双绞线是另一种流行的介质，因为它具有极好的性价比，且具有容易装配的特点。无线信道常用在不可能铺设电缆线路情况下。例如，如果信道穿过人烟稀少的区域或者移动网络用户要进行通信，这时使用无线信道。确保移动性已经影响了电话网络。计算机网络在这方面也无可匹及。虽然如此，在无线技术上架设计算机网络，譬如无线电以太网，被视做无线电通信领域最有前途的技术。有关无线链路的更多细节参见第10章。

8.2.3 传输设备

如图8-1所示，传输线路除了传输介质以外，还包括特定的设备。甚至当传输线路并没有通过传输网络而只是建立在电缆的基础上时，传输线路由**数据电路终端设备**（data circuit-terminating equipment, DCE）构成。

DCE在计算机网络中直接将交换机连接到通信链路。在传统的情况下，DCE包含在通信链路中。DCE设备包括**调制解调器**（modem）（用于电话线路），**ISDN网络终端适配器**（terminal adapters in ISDN network），和用于连接传输网络数据服务单元/电路服务单元（DSU/CSU）的**数字链路连接设备**（devices for connecting to the digital link）。

DCE运作在OSI模型的物理层上，负责从物理介质上接收或发送规定波形、能量和频率的信号。

那些使用通信链路产生传输信息并直接连接到DCE的用户设备有一个通用名称——**数据终端设备**（data terminal equipment, DTE）。计算机、交换机或者路由器都是DTE的例子。这些设备都不包含在通信链路中。

说明　不可能总能够严格界定LAN中的DTE和DCE。例如，LAN适配器可以认为既是计算机（如DTE）的一部分，又是通信链路（如DCE）的一部分。更精确地说，网络适配器的一部分承担DTE的功能；它的另一部分（直接从线路上接收和传输信号部分）是DCE。

有许多连接DCE设备到DTE设备（如计算机、交换机或者路由器）[⊖]的标准接口。它们都在短距离时起作用，通常是几米。

中间设备（Intermediate equipment）通常使用在长距离的通信链路中。主要承担以下功能：

- 改善信号质量
- 在网络的两个用户间创建永久的通信电路

⊖ DTE-DCE接口由V ITU-T系列标准和EIA推荐标准（RS）系列描述。这两套标准在很多方面互相拷贝，最常用的标准是RS-232，RS-530，V.35和HSSI。

在LAN中，可能没有中间设备，如果物理介质的长度——电缆或者无线电波——能够允许一个适配器不用放大信号就从另一个适配器上接收到信号。若不是在这种情况下，需使用中继设备，如转发器（*repeaters*）或者集中器（*concentrators*）（集线器）。

在WAN中，必须确保跨越数百甚至上千英里的高质量信号传输。因此，不用**放大器**（amplifier）（增加信号能量）和**再生器**（regenerator）（放大和重储因长距离传输而失真的信号波形）而要建立长距离的传输线路是无法实现的。

在传输网络中，除了上述考虑到的确保高质量信号的传输设备，也需要转发交换设备——**多路复用器**（multiplexer），**解多路复用器**（demultiplexer）和**交换机**（switche）。这些设备在网络用户间由物理介质（带放大器电缆）创建连续的电路。

依据中继设备的不同，所有的通信链路分为模拟链路和数字链路，在**模拟链路**（analog link）（或者**模拟信道**（analog channel））中，中间设备用来放大模拟信号（例如，拥有连续值的信号）。这些链路通常用在为互联电话交换机的电话网络中。为了创建高速信道，常使用频分复用技术（*frequency division multiplexing, FDM*），复用多个低速本地回路。

在**数字线路**（digital line）中，被传输的信号具有有限的状态值。通常，基本信号——例如，传输设备每个时钟周期发送一次信号——拥有两个，三个或四个状态值。这些状态通过通信线路以方形脉冲或电位方式传输。计算机数据、数字化音频和视频都以此种信号形式传输。事实上，传输网络能成为可能得益于当代计算机、电话和电视网络统一的离散形式的信号表示。在数字通信链路中使用中间设备。重发器改善脉冲的波形，并保证它们同步（例如保存脉冲中间段）。传输网络的中间多路复用和交换设备通过**时分复用**（the time division multiplexing, TDM）规则实现。

8.3 传输链路特性

8.3.1 通信链路中信号的频谱分析

传输线路中传输信号的频谱分布在决定该线路的参数时起了重要作用。根据谐波分析理论，任何周期的过程都可以用不同频率和振幅的正弦振动表示（图8-3）。

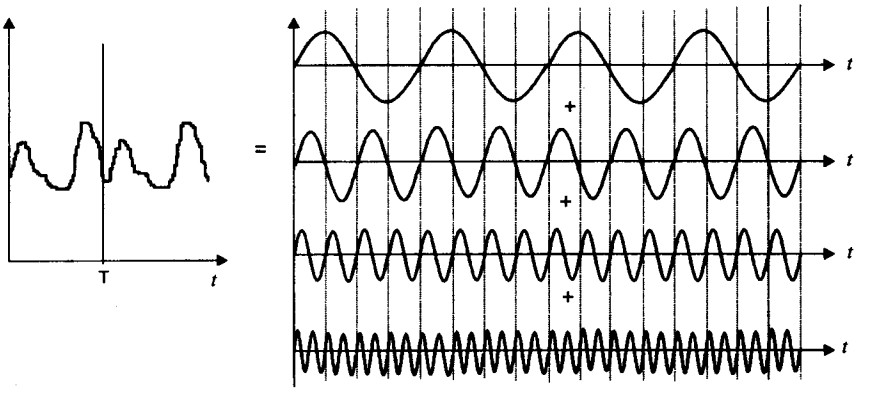


图8-3 周期信号由正弦曲线的和表示

每个正弦曲线被称为**谐波**（harmonic），谐波的整个集合称为**信号频谱**（signal spectrum）。**频谱带宽**（spectrum width）在谐波集合的最大和最小频率解释有所不同，其中谐波的和构成了原信号。

非周期信号可以用连续频率频谱的正弦信号表示。特别是，理想脉冲频谱（单位能量和零持续）包含了整个频率范围的频谱部分，从$-\infty$到$+\infty$（图8-4）。

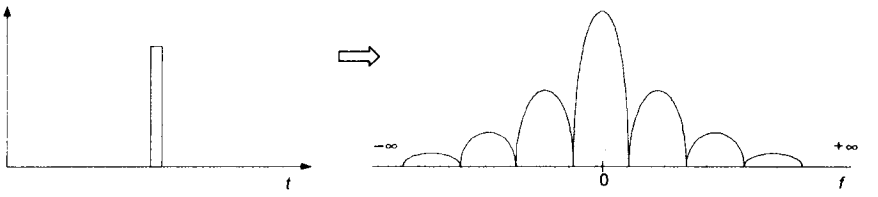


图8-4　理想脉冲频谱

寻找源信号的频谱分析方法众所周知。对于一些可以解析性描述的信号（例如，对于拥有相同持续时间和放大率的方波脉冲序列），频谱可以由**傅里叶公式**（Fourier formula）简单地算出。

对于波形模糊的信号，频谱可以借助于称做频谱分析器的特殊设备得到，分析器测量真实信号的频谱，在屏幕上显示组成谐波的振幅、打印出来、发送用于处理、并存储在计算机中。

经传输线路失真的所有频率的正弦曲线改变了传输信号的振幅和波形类型。当不同频率的正弦波形存在不同程度的失真时，就会导致波形变形。如果这时传输的是模拟信号声音，声音会因为泛音和边缘频率失真而改变音质。当传输脉冲信号时，由于计算机网络的特点，低频和高频谐波会发生失真，因此，脉冲失去方波特性（图8-5）。因此，信号到达线路的另一端却无法识别。

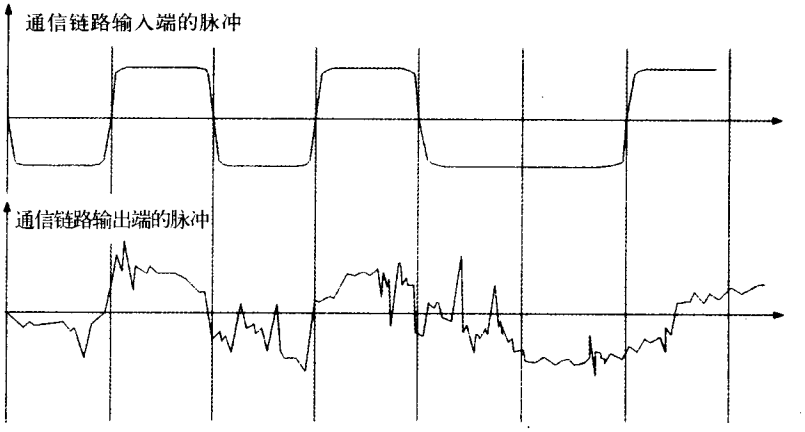


图8-5　通信链路中的脉冲失真

传输线路导致传输信号失真，因为它们的物理参数与理想状态不同。理想传输介质不会将失真引入传输信号，它至少需要零阻抗、电容和自感应。例如，铜线通常由电阻、电容和电感复合作用（图8-6）。因此，对于不同频率的正弦曲线，链路有不同的阻抗值。所以它们的传输也不同。

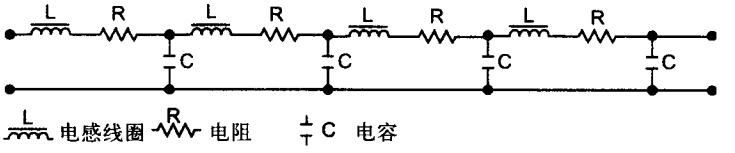


图8-6　作为分布电感和电容负载的通信链路表示

除了传输链路内部的物理参数引起的信号失真外，还有**外部噪声**（external noise）也影响了传输链路输出信号的波形。噪声由各种电子机械、电子设备和大气现象等引起。虽然电缆的设计者采用保护方法，并且放大器和再生器也起作用，但还是无法完全补偿外部噪声带来的影响。除了外部噪声外，还有被称为串扰的**内部噪声**（internal noise）。这种噪声由一对电线产生影响另一对电线。因而链路输出的信号通常具有复杂的波形（如图8-5所示）。

8.3.2 衰减与阻抗

传输链路上的正弦曲线失真程度用衰减和阻抗特性来衡量。

衰减（Attenuation）表明输出通信链路上的参考正弦信号与提供给输入通信链路上的参考信号相比，能量是怎样减少的。衰减（A）通常以分贝来衡量，并用如下公式计算得到：

$$A = 10\lg P_{out}/P_{in} \tag{8.1}$$

这里，P_{out}是输出线路上的信号能量，而P_{in}是提供给输入线路上的信号能量。因为衰减取决于通信链路的长度，因此称做**线性衰减**（linear attenuation），并作为链路特性（例如，特殊长度通信线路衰减）。对于LAN电缆，一般为100m的链路。这一值是局域网技术所能达到的最大电缆长度。对于WAN链路，线性衰减的长度是1km。

通常，衰减描绘了由电缆和交叉连接组成的并且不经过放大器和再生器的通信链路的无源部分。因为电缆没有经过中间放大器的输出信号的能量总是比输入信号的能量低，所以电缆衰减总是负值。

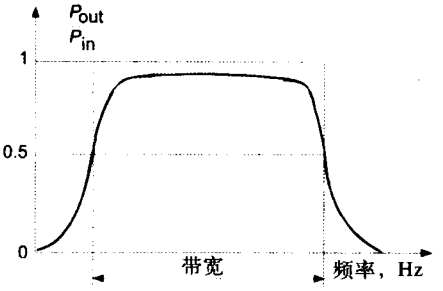


图8-7 衰减频率的相关性

正弦信号的能量衰减程度取决于正弦波的频率；这种相关性也用来刻画通信链路（图8-7）。

通常，当描述通信链路的参数时，只提供某些频率（*for only some frequency*）的衰减值。只使用部分值而不是全部属性，一方面，与测试链路质量时的方法有关，另一方面，与传输信号的主频有关，主频通常事先就已经知道了。这一频率是谐波拥有最大振幅和能量时的频率。因此，知道了这些频率的衰减，对于粗略评估使用该线路传输信号时的失真变形已经足够了。

说明 如上所述，电缆衰减总是负值。尽管如此，专家在谈论衰减的时候常常略去负号。他们可能会说线路信号越好，衰减越小。这一论述只有指衰减的绝对值时才是正确的。如果考虑符号，那么就是衰减越大，通信链路的质量越好。考虑一个例子，室内使用5号双绞线，所有的LAN中都使用该种电缆。对于100MHz的频率衰减不小于−23dB，最大长度100m。对于更好质量的6号电缆，在100MHz时，衰减不低于−20.6dB。当然，−20.6>−23.6，同时，20.6<23.6。

对于非屏蔽双绞线（5号线和6号线）衰减和频率的依赖关系如图8-8所示。

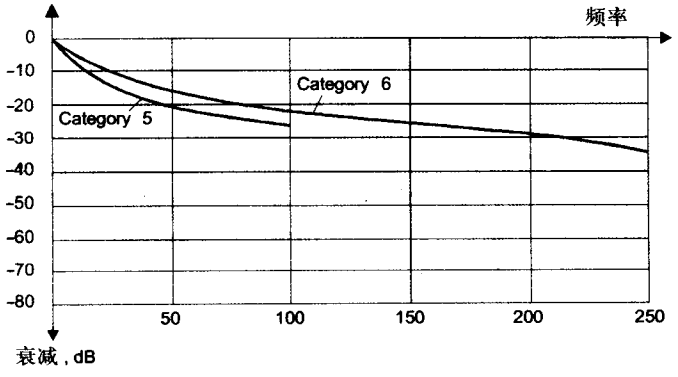


图8-8 非屏蔽双绞线电缆的衰减

光纤电缆衰减系数显著减少（取绝对值），对1 000m长的电缆通常−2.0到−3.0。因此，它的质量比双绞线好很多。实际上，所有的光纤都有复杂的衰减与波长的依赖关系，有三个称为**透明窗口**

(transparency window)。图8-9展示了光纤的典型衰减相关性。从图中清楚地发现现代光缆的有效应用区域限制在波长为850nm，1300nm和1550nm（相应的频率为35THz，23THz和19.4THz）。1550nm的窗口确保最少的丢失，并且对于具有固定能量的发送端和接收端灵敏度有最远的传输距离。

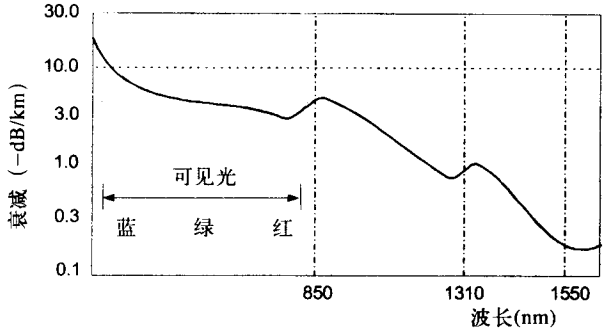


图8-9　光纤的透明窗口

绝对能量级和相对能量级用来描述信号能量。**绝对能量级**（absolute power level）以瓦特来计量。**相对能量级**（relative power level）与衰减相似，以分贝计量。同时，在度量信号能量时，1mW用作参考值。因此，相对能量级p用以下公式计算得到：

$$p = 10 \lg P/1\text{mW [dBm]} \tag{8.2}$$

这里，p是以毫瓦为单位的绝对信号能量，dBm是相对能量级的度量单元（分贝每毫瓦）。

当计算传输链路的能量预算时，使用相对能量值非常方便。

示例　假设我们要计算发送端的最小相对能量x（dBm），为确保输出信号的相对能量不低于阈值（dBm）。已知线性衰减等于A。假设X和Y分别是线路输入和输出的绝对值，以mW为单位。

定义：

$$A = 10 \lg X/Y$$

运用对数属性：

$$A = 10\lg X - 10 \lg Y = 10\lg X/1\text{mW} - 10\lg Y/1\text{mW}$$

注意，等式的最后两项为输入和输出线路上的信号相对能量。得到下列简单关系：

$$A=x-y$$

由此，发送端的最小能量可以用衰减和输出信号的能量的和来定义：

$$x=A+y$$

这个计算相当简单，因为把输入和输出信号的相对能量作为初始数据。

y值称为**接收敏感度阈值**（receiver sensitivity threshold）。它是在接收端的输入端能够正确识别出信号中所包含的离散信息的最小信号能量。对于通信链路的准确操作，必须保证传输信号的最小能量超过接收端的灵敏度阈值：$x-A>y$。传输链路的能量预算背后的主要思想即是这一条件的审核。

铜质通信链路的另一个重要参数是**阻抗**（impedance）。这一参数完全阻止特定频率的电磁波在均匀电路中传播。阻抗用欧姆度量，并且取决于线路参数诸如正电阻、线性自感、线性电容以及信号频率。发送端的输出电阻与线性阻抗相一致。否则，信号衰减将会太多。

8.3.3　抗噪声与传输可靠性

线路抗噪声（Noise immunity of the line），如同它的名字一样清楚，它决定了线路所能承受

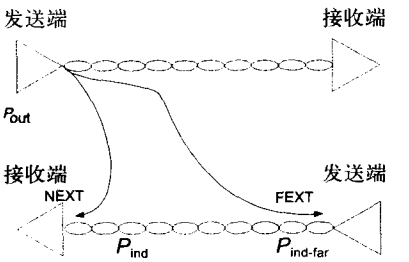
的由周围环境或者是电缆本身的内部导体产生的噪声影响。线路抗噪声依赖于所使用的物理介质、屏蔽层和线路的防噪属性。无线电信道的抗噪性最差。铜质线路有稍好的防护能力，而光纤电缆具有最好的抗噪声，它能阻止外部电磁无线电波。通常，为了减少外部电磁场的噪声，采用屏蔽或者双绞线。

电子耦合和电磁耦合也是影响铜质电缆噪声干扰的因素。**电子耦合**（electric coupling）由被作用电路上的电流和作用电路上的电压的比值决定。**电磁耦合**（magnetic coupling）则是被作用电路上的电动势和作用电路上的电流的比值。电子耦合和电磁耦合的结果即是被作用电路上的串扰信号。有多个参数用于描述针对串扰的电缆稳定性。

近端串扰（near end cross talk, NEXT）决定了电缆的稳定性，如果串扰是由连到相邻双绞线的发送端产生的信号影响。这种情况下，电缆的接收端连接到被作用双绞线（图8-10）。NEXT用分贝表示，为$10\lg P_{out}/P_{ind}$，这里P_{out}是输出信号能量，P_{ind}是感应信号能量。

NEXT值越小（考虑符号），电缆越好。例如，对于5号双绞线，在100MHz时NEXT值必须小于−27dB。

当发送端和接收端连接到电缆的不同端（不同线）时，**远端串扰**（far end cross talk, FEXT）评估电缆的稳定性。线路的这一参数通常好于NEXT，因为信号到达电缆的远端后，每组线都衰减。



P_{out}—发送端输出端信号的能量
P_{ind}—电缆近端的感应信号的能量
$P_{ind\text{-}far}$—电缆远端的感应信号的能量

图8-10 瞬时衰减

通常，NEXT和FEXT用于由多种绞线组成的电缆，因为相互串扰可能达到相当大的值。对于单线同轴电缆（例如，由单股屏蔽线构成），这个参数没有任何意义。对于双股同轴电缆，也不使用该参数，因为每股都采取了高度保护。光缆之间也不产生任何可察觉的噪声。

新网络技术的应用使得同步数据传输可以通过使用多个双绞线实现，串扰参数又由PowerSUM（PS）前缀，如PS-NEXT和PS-FEXT，引入。这些参数反映了电缆的稳定性与影响单个电缆对和组成电缆的剩余对的串扰的总能量的关系（图8-11）。

电缆保护（cable protection）表明**串扰率衰减**（attenuation to cross-talk ratio, ACR）是电缆的另一个重要特性。保护定义为有效信号和噪声之间的级差。ACR值越高，则根据香农定律求得电缆上可以传输的潜在数据率也越高。图8-12的特性反映了非屏蔽双绞线电缆ACR值对单个频率的相互关系。

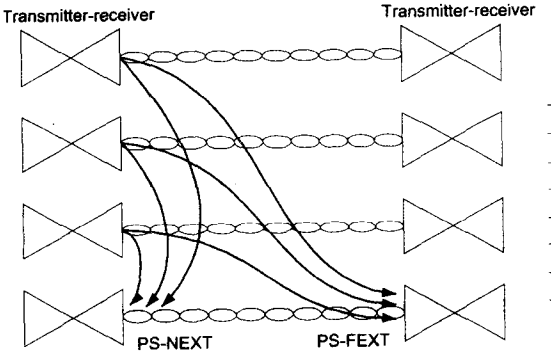


图8-11 总传输衰减

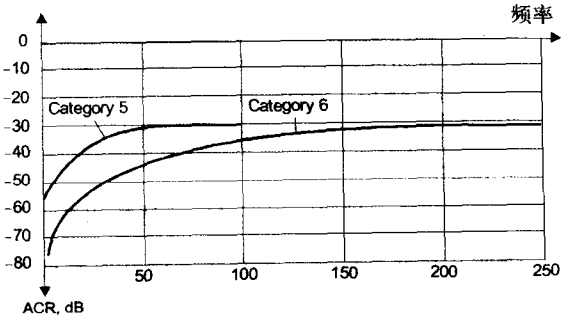


图8-12 非屏蔽双绞线的ACR值

数据传输可靠性（data transmission reliability）描述了传输数据的任意位发生失真的可能性。

有时这一参数称为**位错误率**（bit error rate, BER）。通常，没有额外纠错设备（如自纠错编码或支持失真帧重传的协议）的通信链路的BER值在10^{-4}~10^{-9}间。在光纤线路中，这一值为10^{-9}。数据传输可靠性值为10^{-4}，位于平均水平，也就是说传输10 000位会有1位失真。

8.3.4　带宽与容量

带宽（bandwidth）是指没有超出线路预先设定范围限制的连续频率段。这意味着带宽决定了正弦信号的频率范围。在这一范围内，信号在线路上的传输不产生明显失真。

通常，频率限制被为输出信号能量与输入信号相比减少到原来一半的频率，对应于衰减为−3dB。如同稍后要解释的一样，带宽对使用通信链路的最大传输数据率有最大影响。

带宽依赖线路的类型和长度。图8-13表明了常用通信链路的带宽和常用技术频率范围。

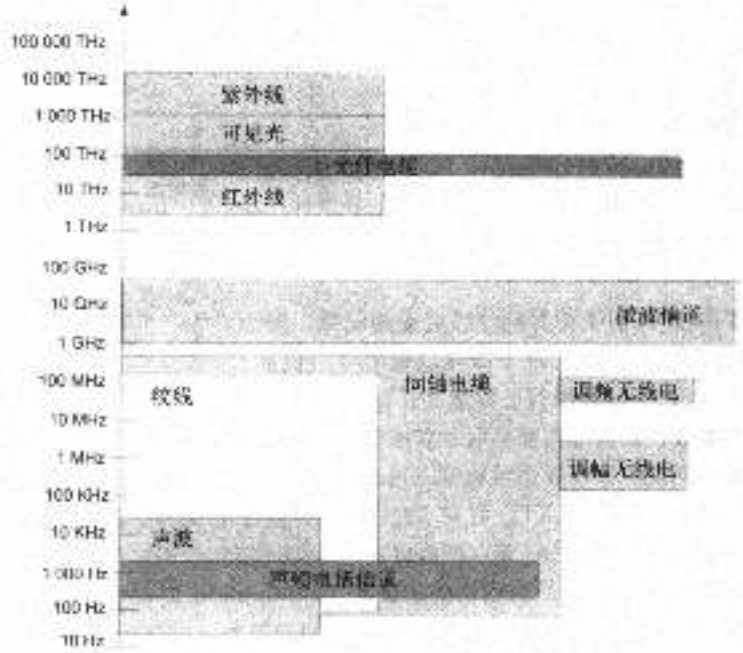


图8-13　常用通信链路的带宽和频率范围

链路容量（link capacity）是指这一线路可以到达的最大数据率。这一属性依赖物理介质（*physical medium*）的参数和数据传输方法（*data transmission method*）。因此，在定义物理层协议之前不可能讨论线路容量。

例如，因为数据链路上限定数据比特率的物理层协议已经定义，那么数据链路层的吞吐量也预先知道——64Kb/s，2Mb/s等。

当必须确定用在特殊链路上的已有协议时，其他链路属性——如带宽，串扰参数和抗噪声——变得十分重要。

容量与信息率相似，用每秒的比特数和接收单元如每秒千比特（Kb/s）来衡量。

说明　传统上线路和通信设备的容量和信息率都以每秒比特数来衡量，而不用每秒字节数。这是由于数据在网络上是串行传输的，与计算机内部的并行（以字节）传输不同。在网络技术中，度量单位如千比特，兆比特和吉比特对应于10的幂（这里1Kb是1 000比特，1Mb是1 000 000比特）而不是程序中的2的幂。这里$1KB=2^{10}=1\ 024$，$1MB=2^{20}=1\ 048\ 576$。

链路容量不仅依赖衰减和带宽，还依赖于传输信号的频谱。如果信号的谐波（那些振幅构成主要信号的谐波）落在链路带宽内，那么这一链路传输高质量的信号，且接收端可以识别由发送端通过链路发送来的信息（图8-14a）。如果谐波落在链路带宽限制之外，信号就会明显的失真，并且接收端也会在识别过程中出错（图8-14b）。

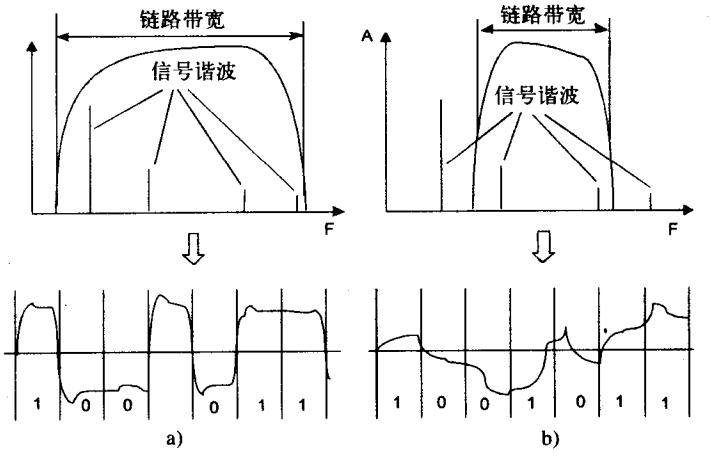


图8-14 链路带宽和信号频谱间的对应关系

8.3.5 比特与波特

通信链路上的离散信息信号的表示方法称为*物理*（*physical*）或者*线路编码*（*line encoding*）。信号频谱和链路吞吐量依赖于选择的编码方式。

所以，对于不同的编码方式，同一链路的容量也是不同的。例如，当使用10Base-T物理层编码方式时，3号双绞线可以以10Mb/s的带宽传输数据。如果使用100Base-T4标准，则链路容量可以达到33Mb/s。

说明 根据信息理论的基本原理，任何接收信号独特和不可预见的改变都承载了某种信息。因此，以确定的振幅、相位和频率接收的正弦波将不带有任何信息，因为虽然发生了信号改变，但是这些都是可预见的。同样，控制总线上时钟线路上的时钟脉冲也不带有任何的信息，因为他们的改变也是可预见的。数据总线上的脉冲无法事先预测。这一特性使得它们提供资讯，因为这些脉冲在计算机单元及设备间传输信息。

大多数编码方法改变周期信号的特殊参数——正弦曲线频率、振幅，或者相位和脉冲序列的潜在信号。改变参数的周期信号称为**载波信号**（carrier signal）或者**载波频率**（carrier frequency），因此，正弦曲线被用来做这种信号。改变载波信号参数的过程要根据所传输的信息，称为**调制**（modulation）。

如果信号能够只用两个状态来区分，那么这种信号的改变对应于1bit——也是信息的最小单元。如果信号的状态多于两个，那么这一改变就要用*多位信息*（*several bits of information*）。电信网络中离散信息的传输是时钟计时的，也就是说信号在称为**时钟**（clock）的固定时间间隔后改变。这也意味着信息接收端认为每一时钟开始会有新的消息送至输入端。基于这点，接收端从转发器获得新信息，而不管信号为重复上一时钟状态还是与前一时钟状态不同。举例来说，如果时钟的持续时间是0.3s，并且信号有两个状态，电位5V编码为1，那么5V的信号在接收端输入端持续3s，等价于传输用如下二进制表示的信息：1111111111。

载波信号每秒钟信息参数的变化次数用*波特*（*baud*）来衡量。一波特等于信息参数每秒改变

一次。信息信号两次变换之间的时间称为传输时钟。

要点　信息率与波特率一般不相等。可能大于、小于或者等于波特率。它们之间的关系依赖于所选择的编码方式。

如果信号多于两个不同的状态，那么信息率以每秒比特数计就会高于（*higher*）波特率。假设我们选择正弦信号的相位和振幅作为信息参数，那么就有四个可辨别的相位状态——0，90，180和270度——和两个可辨别的信号振幅。这表明信息信号可以有八个不同的状态。因此，信号的每次改变要占用3比特信息。在这种情况下，调制解调器以2 400波特的速率运作（意味着可以每秒改变信息信号2 400次），以7 200b/s的速率传输信息，因为信号的每次改变传输3比特信息。

如果信号只有两种状态（也即承载1比特信息），那么信息率通常和波特率相同。尽管如此，情况可能完全相反：信息率可能低于波特率。这种情况在如下条件下发生，在确保接收端可以可靠识别用户信息时，载波信号信息参数的多次变化使用1比特编码。例如，当脉冲正极用1表示，脉冲负极用0表示，那么物理信号每传输一比特改变两次。在使用这种编码方式时，比特率比波特率低两倍。

载波信号的频率越高，调制频率也越高。因此，可以达到高输出的传输链路。

但是，载波信号的频谱带宽随着频率的增加而增加。线路带失真的频谱由带宽决定。链路带宽和传输信号频谱带宽之间的差异越大，信号就越容易失真，在接收端发生错误的可能性就越大。因此，可能的信息率应该低些。

8.3.6　带宽与容量间的相关性

链路带宽和容量间的关系，与所选择的物理编码方式无关，而由克劳德·香农（*Claude Shannon*）发布：

$$C = F \log_2(1 + P_c / P_{\text{noise}}) \tag{8.3}$$

这里，C是每秒的链路容量，F是单位为赫兹的链路带宽，P_c是信号能量，P_{noise}是噪声能量。

从公式看出，固定带宽的线路没有理论上限。实际中，这一限制是存在的。可以通过增加传输能量和降低线路噪声来增加线路容量。这两个组成部分都难以改变。增加传输能量会显著增加大小和开销。降低噪声等级需要有高质量防护电缆，却非常昂贵。而降低发送端和传输设备中的噪声很难实现。进一步而言，有效信号和噪声的能量对容量的影响局限于对数相关性，其增长比线性相关性慢。这导致了典型的信噪比初始值为100。传输信号能量翻倍线路容量只增加15%。

另一种关系，来自亨利·奈奎斯特（*Harry Nyquist*），与香农公式相似，决定了通信链路（如容量）的最大范围。但没有考虑线路噪声。

$$C = 2 F \log_2 M \tag{8.4}$$

这里，M是信息参数的不同状态数。

如果信号有两个不同的状态，那么范围是线路带宽的两倍（图8-15a）。如果发送端使用可用于数据编码且有多于两个稳定状态的信号，那么线路的范围增加。这是由于发送端每个时钟周期都发送多位数据——例如，2比特，提供四个不同信号状态是可行的（图8-15b）。

虽然奈奎斯特公式没有明显地考虑噪声，但噪声影响隐含反映在信号状态数目的选择上，为了增加链路容量，增加信号状态数是有意义的。在实际中，线路噪声代表了使用该方法的障碍。例如，线路容量（图8-15b）能够再次加倍，使用16个信号等级来替代4个用于编码。但是，如果噪声的振幅刚好超过了相邻层之间的差别，接收端将无法可靠识别出传输数据。因此，信号的可能状态数受到有效信号能量和噪声能量比率的限制，并且由奈奎斯特公式决定了在给定状态数前提下的最大传输率，也将接收端识别稳定信号的可能性考虑进去。

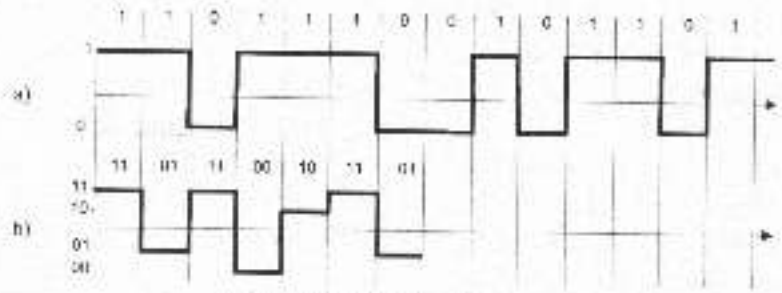


图8-15 通过额外信号线来增加信息率

8.4 电缆类型

当今，三种类型的通信线路在市内和室外都得到应用。

- 双绞线（*Twisted pair*）
- 同轴电缆（*Coaxial cable*）
- 光纤电缆（*Fiber-optical cable*）

8.4.1 非屏蔽和屏蔽双绞线

一对缠绕在一起的电线称做双绞线（*twisted pair*）。这种类型的介质非常流行并且在内部和外部形成了大多数电缆的基础。一条电缆可能包含多条双绞线。外部电缆有时由几十条双绞线构成。

缠绕电缆就减低了外部噪声的影响和通过电缆传输的有效信号串扰。

电缆结构的主要特点由图8-16简要列出。

基于双绞线的电缆称做对称（*symmetric*）电缆，由同样结构的两条电线组成。对称电缆可能是屏蔽的，基于STP，也可能是不屏蔽的，基于UTP。

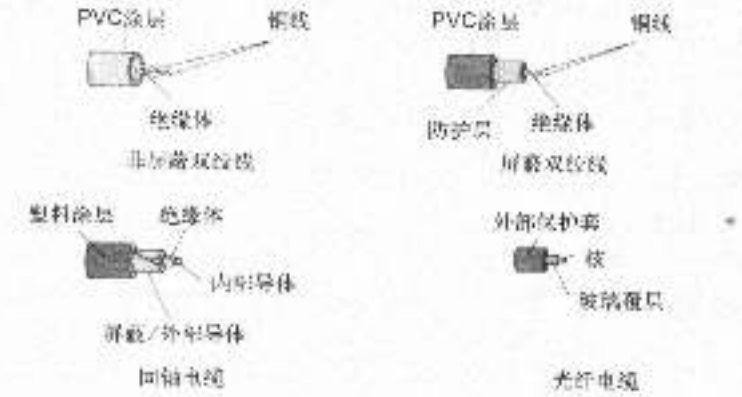


图8-16 电缆设计

有必要区分导线的电子绝缘层（*electric insulation*），其存在于任何电缆中，和电磁屏蔽层（*electromagnetic shielding*）。电子绝缘层由绝缘纸层或聚合层组成，例如聚氯乙烯树脂（PVC）或者聚苯乙烯。除了电子绝缘层，导体还由电磁屏蔽层包裹，大多由铜质套管组成。

用于室内的铜质非屏蔽双绞线电缆根据国际标准分为多种（1类——7类）。

- 1类电缆（category 1）。当传输率要求最低时使用。通常，用于数字或者模拟声音的传输，也适用于低速（上限20Kb/s）数据传输。直到1983年，这类电缆还是电话电缆的主要适用类型。

- **2类电缆**（category 2），最初由IBM公司在建立他们的专用电缆系统时使用。对这种电缆的主要需求是在传输信号频谱的上限为1MHz时变得尤为突出。
- **3类电缆**（category 3），随着商用电信电缆系统标准（EIA-568）的发展，该标准在1991年被标准化。在EIA-568标准下发展出现了EIA-568A标准。EIA-568标准定义了3类电缆的电子属性，包括把频率范围增大到16MHz。因此，3类电缆能够支持高速网络应用。
- **4类电缆**（category 4），改进了3类电缆，4类电缆必须通过频率等于20MHz的信号传输测试，并确保改善噪声防护和低信号损失。这类电缆在实际中很少使用。
- **5类电缆**（category 5），专门设计用来支持高速协议。因此，它的参数上升到100MHz。大多数高性能标准致力于使用5类双绞线。这一电缆与高速协议一起使用，数据传输率在100Mb/s，比如FDDI，高速以太网和更快的协议——速率在155Mb/s的ATM和速率在1 000Mb/s的吉比特以太网。5类电缆已经替代了3类电缆，目前这种电缆（和光纤电缆一起）用在大型建筑间建立新的电缆系统。
- **6类和7类电缆**（category 6&7），到最近才开始工业化生产。对6类光缆，定义上限频率为250MHz。对7类电缆，频率达到600MHz。7类电缆必须是屏蔽的，可以应用到单对和整个电缆。6类电缆可以是屏蔽或者非屏蔽的。这些电缆比起5类UTP电缆来说，更趋向于在长距离上支持高速协议。

所有类型的UTP电缆按照4对电缆生产。每对拥有独特的颜色和电线层。通常两对用于数据传输，另两对用于声音传输。

屏蔽双绞线（*shielding twisted pair*）很好地保护传输信号不受外部噪声干扰。除了这点，还具有较小的电磁波辐射，因此可以保护网络用户不受有害电磁辐射的影响。但是，接地防护层提高了电缆的价格，并使得电缆的安装过程更加复杂，因为需要高质量的接地。

IBM专用标准是决定趋向应用于室内的屏蔽双绞线参数的主要标准。根据这一标准，电缆被按照类型而不是分类分为：类型1，类型2，……，类型9。

类型1（*type* 1）电缆根据IBM标准由两对有传导电缆层的缠绕线组成，其中电缆层接地。类型1电缆参数与5类UTP电缆大致相同。但是，类型1电缆的阻抗为150欧姆，比5类UTP电缆（100欧姆）高很多。因此，通过简单地用STP类型替换非屏蔽UTP来改进电缆防护层是可行的。发送端倾向于使用100欧姆阻抗的电缆而不会满足于150欧姆阻抗的电缆。

8.4.2 同轴电缆

同轴电缆（coaxial cable）由*非对称导体对*组成。每对是内部铜线和外部轴线，可以是空铜管或者由与内线分离的电绝缘介质。外部线扮演了两种角色。首先，它用于传输信息信号；其次，它是保护内部线不受外部电磁域影响的防护层。有多种同轴电缆，它们的特性和应用领域各不相同，可以应用于局域网、广域网、电信传输网和电缆电视等等。

目前的标准认为同轴电缆不是构建电缆系统的最佳选择。如下是根据美国分类标准的主要类型和这些电缆的特点。

- **粗同轴电缆**（“thick” coaxial cable）为以太网10Base-5网络开发。阻抗50欧姆且外径为0.5英寸（约合12mm）。这类电缆含一个厚的内部导体（直径2.17mm），确保很好的电子和机械属性（在10MHz处衰减，并不会比18dB/km更差）。但这类电缆不够柔软，因此不易安装。
- **细同轴电缆**（“thin” coaxial cable）用于以太网10Base-2网络。阻抗50欧姆，但是与粗同轴电缆相比，机械和电子属性明显相差很多。细同轴电缆的直径0.89mm，但柔软，方便安装。与粗同轴电缆相比衰减明显升高，导致必须减少电缆长度以便于在同一段中获得同样的衰减。
- **电视电缆**（the TV cable）的阻抗为75欧姆。被广泛用在电缆电视中。有用该电缆传输数据

的局域网标准。

8.4.3 光缆

光纤（optical fiber）由细（5~60μm）柔软玻璃纤维（光波导）组成，光信号可沿着玻璃纤维传播。这种电缆具有最好的质量，因为它可以确保以非常高的速率（10Gb/s或更高）传输数据。此外，与其他任何传输介质相比，可以更好地保护数据不受外部噪声的干扰。因为光传播的特殊属性，很容易对信号加以保护。

每一光波导由中心波导（核）——光纤维——和与核相比有更小折射率的玻璃层组成。光波通过核传播而不会偏离，因为所有光都从外层反射。根据折射因子和核直径，光纤分为以下几种：

- 折射因子逐层改变的多模式纤维（MMF）（图8-17a）
- 折射因子光滑改变的MMF（图8-17b）
- 单模式纤维（SMF）（图8-17c）

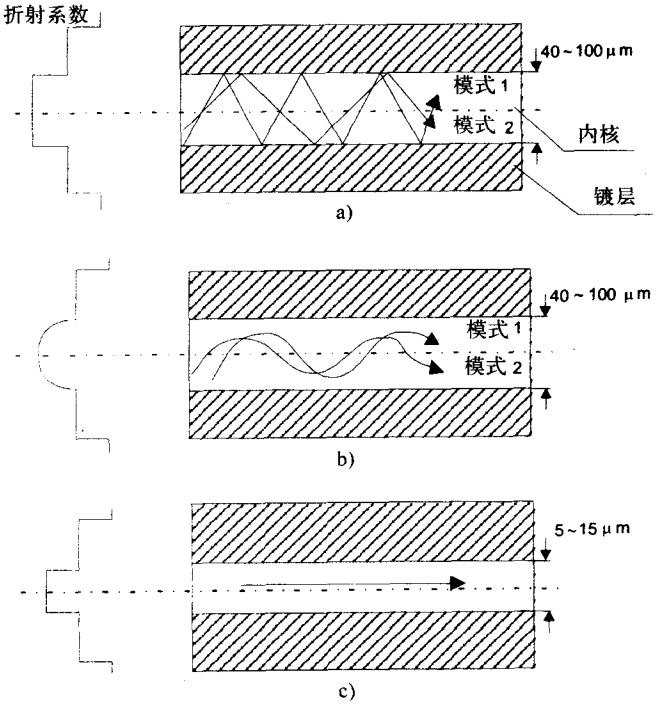


图8-17 光纤电缆类型

模式的概念描述了在电缆内核中光线的传播模式。在SMF中，中心核的直径从5μm到10μm，与光波长相比非常小。在这些条件下，所有的光波沿着波导光轴传播时，实际并不从外层反射。为*SMF*生产高细、高质量的纤维是一个复杂的过程。因此，SMF光缆十分昂贵。而且，实现毫无能量损失地将光束导入如此小直径的纤维中也是困难的。

在MMF电缆中，使用稍宽的内核，更易于生产。在MMF中，核同时传导多条光线。这些光线以不同的角度从外层反射。光线的反射角称为**光线模式**（the ray mode）。在折射因子光滑改变的多模式电缆中，传播模式的属性相当复杂。不同模式的光线干涉降低了传输信号的质量，导致了传输脉冲的失真。MMF电缆易于生产，因此要比单模式电缆廉价很多。同时，它的属性也比单模式电缆差很多。

因此，多模式电缆主要用于以不高于1Gb/s速率短距离（300~2 000m）传输数据，而单模式电

缆趋向于以最高速率　每秒几十占比特的速率（使用DWDM技术甚至可以达到数仗比特每秒的速率）传输数据，并跨越从数公里（局域网或城域网）到几十甚至数百公里（长距离通信）的范围。

作为光源，光纤电缆使用：

- 发光二极管（LED）
- 激光二极管

对于单模式电缆，只使用激光二极管；光纤直径很小，LED产生的光束因其很宽的直射图样，无法在没有显著损失的情况下导入波导核。与LED相比，激光二极管拥有狭窄的直射图。基于这一点，较廉价的LED只用于MMF电缆。

光纤的成本并没有明显超过基于双绞线的电缆。但在处理光纤电缆时，安装工作非常困难并且昂贵，这是因为所有安装操作的高劳动强度和安装设备的昂贵。

8.4.4　楼宇的结构化布线系统

结构化布线系统（structured cabling system, SCS）是交换元素（电缆、连接器、插槽、交换板和储藏室）和共享使用方法的集合。它在计算机网络间创建规则和易扩展的通信结构。建筑本身就是一个规则结构，它由楼层组成，并且每个楼层包含一定数量的由通道连接的房间。建筑结构定义了布线系统的结构。

楼宇的结构化布线系统是一种建筑块的集合，网络设计者通过标准电缆连接连接器和交换板，构建需要的配置。必要时，连接的配置可以简单地更改，例如，添加计算机、段和交换机，移除不必要的设备和改变计算机与集线器间的连接都是可行的。

如今，已经很好地定义了用于商业建筑的布线系统。创建布线系统过程的层次方法称做结构化。在商业建筑SCS的基础上，隶属于不同组织或者同一组织的不同部门之间的局域网变得可操作。计划SCS，并层次性地建立在主干和多个T型接头上（图8-18）。

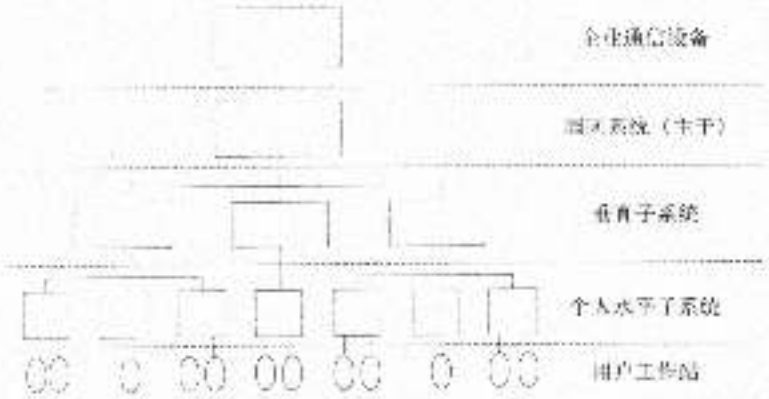


图8-18　结构化布线系统层次子系统

典型的SCS层次结构（图8-19）包括：

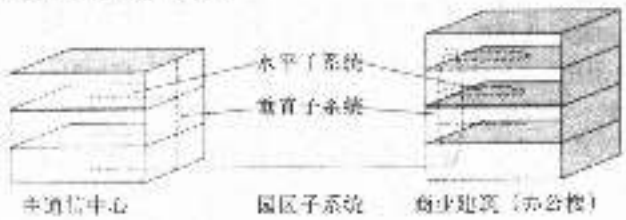


图8-19　电缆子系统结构

- **水平子系统**（horizontal subsystem）与楼层关联：它们用终端用户插口连接楼层交换室。
- **垂直子系统**（vertical subsystem）把每一层的交换室连接至整个建筑的中央设备室。
- **园区子系统**（campus subsystem）把数个建筑连接到整个园区的中央设备室。布线系统的这一部分通常称做主干。

在使用结构化布线系统时，以随机方式铺设电缆为企业提供了许多好处。如果SCS系统经过周密计划，那么它可以提供通用介质（*universal medium*）用来在局域网中传输数据，组织局部电话网络，传输视频信息甚至从防火墙安全系统传感器发送信号。这也允许大多数控制、监控和管理不同企业服务过程的自动化，如家庭和生命支持系统。

更进一步，SCS的使用让添加用户或更换工作站的过程更加有效和经济。众所周知，布线系统的成本主要取决于它的安装成本而不是电缆的成本。因此，事先完成电缆安装更加有效，可能在安全环节有冗余。这要比简单扩展长度多次改变电缆布局好得多。

小结

- 根据传输设备的类型，所有的通信链路分为模拟和数字。在模拟链路（模拟信道）中，中继设备用来放大模拟信号，模拟信道使用频分复用（FDM）和波分复用（WDM）。
- 在数字链路中，信号以确定数目的状态传输。数字信道使用特殊的中继设备——再生器，用于改善脉冲波型以确保重新同步（例如存储脉冲间隔）。通过时分复用（TDM）原理实现传输网络的中间复用和设备交换，在这一原理下，每个低速链路都可以在高速信道上获得专属的时隙。
- 带宽描述了通信链路在可接收衰减下传输频率的范围。
- 链路容量依赖于内部参数，尤其是带宽；依赖于外部参数，例如噪声等级和噪声抑制程度；依赖于所采用的离散数据编码方法。
- 香农公式定义了在固定链路带宽和信噪比下的通信链路的容量。
- 奈奎斯特公式根据带宽和信号状态数表示链路容量。
- 基于双绞线的电缆分为非屏蔽（UTP）和屏蔽（STP）。UTP电缆易于生产和安装，但STP可以提供高质量的保护。
- 光纤电缆有杰出的电磁和机械特性。但缺点是复杂而昂贵的安装。
- 结构化布线系统是一套通信元素——电缆、连接器、插槽、交换板和交换室——符合标准且允许建立规则和可扩展的通信结构。

复习题

1. 传输链路和电路之间的区别是什么？
2. 电路能包括链路吗？链路能包括电路吗？
3. 数字信道能传输模拟数据吗？
4. DTE和DCE的功能是什么？网络适配器属于哪种设备？
5. 下面分别属于哪种通信链路特性类型：噪声等级、带宽和线性电容？
6. 用什么方法可以增加链路信息率？
7. 为什么不能总是通过增加信息信号状态数来增加信道容量？
8. 什么机制用来抑制UTP电缆的噪声？
9. 哪种电缆保证更高的信号传输质量：高NEXT参数值电缆，低NEXT参数值电缆？
10. 理想脉冲的频谱宽度是什么？
11. 列出光纤电缆的类型。

12. 如果用UTP电缆替代STP电缆会发生什么变化？
13. 列出结构化布线系统的主要优点。
14. SCS水平子系统使用何种电缆？
15. 在水平子系统中使用光纤电缆有什么问题？

练习题

1. 给定下列值：
   - 最小传输能量P_{out}（dBm）
   - 电缆衰减A（dB/km）
   - 接收端敏感阈值P_{in}（dBm）

   求出信号可以完全传输的最大通信链路的长度。
2. 信道带宽为20kHz，发送端能量为0.01mW，信道的噪声级别为0.0001mW，那么使用信道的传输率（b/s）理论上限值为多少？
3. 带宽为600kHz，使用10状态信息信号编码方式，计算全双工通信链路的容量。
4. 计算传输128字节数据包（信号传播速率可认定为真空中光传播速率——300 000 km/s）的信号传播延迟和信号传输延迟：
   - 传输率为100Mb/s的100m双绞线电缆
   - 传输率为10Mb/s的2km同轴电缆
   - 传输率为128Mb/s的72 000km卫星信道
5. 给出传输时钟频率为125MHz，信号有5个状态，计算信道率。
6. 发送和接收网络适配器连接到相邻的UTP电缆对。如果传输能量为30dBm，且电缆的NEXT参数值为20dB，那么接收端输入端的感应噪声能量为多少？
7. 如果调制解调器以33.6Kb/s双工方式传输数据，计算如果信道带宽为3.43kHz，信号有多少个状态？

第9章 数据编码和多路复用

9.1 引言

前一章考虑的导向介质只是提供了传输离散信息的潜在可能性。为了能够让发送端和接收端通过特定介质连接完成信息交换，必须在传输离散信息时，对表示与信号相关的二进制1和二进制0达成一致。两种信号用来在传输介质中代表离散信号：方脉冲和正弦波。在第一种情况下，使用编码方法代表离散信号，在第二种情况下，使用调制方法。

根据在相同的信息速率条件下信号频谱带宽的不同，有多种编码方法。为了能够保证信息传输的最小错误率，链路带宽应比信号频谱宽。如果不能满足此条件，用于代表1和0的信号将明显失真，接收端将无法正确识别出传输的信息。因此，信号频谱是衡量编码方法有效性的重要指标之一。除了这些以外，编码方法还应该使接收端和发送端同步且保证一个可以接受的信噪比。这些需求相互冲突。因此，所使用的每种编码方法都是在考虑主要需求后的一种折中。

即使所选择的编码可以提供好的同步水平和高值的信噪比通信链路中的比特错误不能被完全消除。因此，传输离散信息时，使用特殊的编码。这些编码可以侦测比特错误，有些编码甚至可以纠正错误。

本章最后还介绍了复用技术，复用技术提供了在单条传输链路上创建多条信道的可能。

9.2 调制

9.2.1 传输模拟信号时的调制

从历史的角度，调制首先用于传输**模拟**信号而不是**离散**信号。

当在落入频谱高频范围内的信道上传输低频模拟信号时，需要调制模拟信号。这种情况的例子包括用无线电和TV传输声音。人类声音的频谱宽度大约在10kHz左右，无线电范围使用更高一些的频率——从30kHz到300MHz。TV使用甚至更高的频率。显然通过该介质直接传输声音是不可行的。

因此，为了解决这一问题，决定根据低频信号的变化改变（调制）高频信号的振幅（图9-1）。在这种情况下，结果信号的频谱嵌在高频范围内。这种调制称为**调幅**（amplitude modulation, AM），因为高频信号的振幅是承载信息的参数。

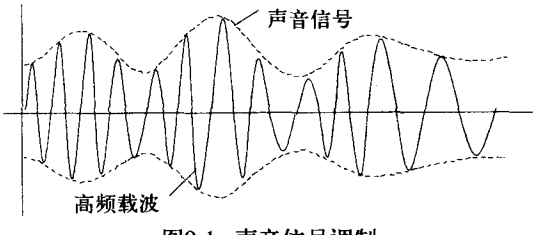


图9-1 声音信号调制

除了载波振幅，频率也可以用来作为信息参数。在这些情况下，我们采用**调频**（frequency modulation, FM）[⊖]。

信号的高频部分也称做*载波频率*（*carrier frequency*），因为这一信号起到了低频信息信号载波的作用。

9.2.2 传输离散信号时的调制

当用调制传输离散信息时，根据载波正弦信号的振幅、频率和相位变化编码1和0。在调制信号传输离散信息时，使用如下方法：

⊖ 注意当调制模拟信息时，不使用相位作为信息参数。

- 幅移键控（ASK）
- 频移键控（FSK）
- 相移键控（PSK）

或许在使用调制传输离散信息时，已知的最好例子就是用电话线传输计算机信号。标准用户线路也称做声道（tone channel），它的典型幅频特性如图9-2所示。电路通过电话网络的转接交换机连接用户的电话机。声道传输频率范围从300到3 400Hz，因此它的带宽为3 100Hz。如此狭窄的带宽对于高质量声音传输已绰绰有余，但它对于传输方脉冲形式的计算机数据却不够。解决这一问题的方法就是幅移键控。在发送端调制载波正弦波和在接收端对它解调的设备称做**调制解调器**（modem, modulator-demodulator）。

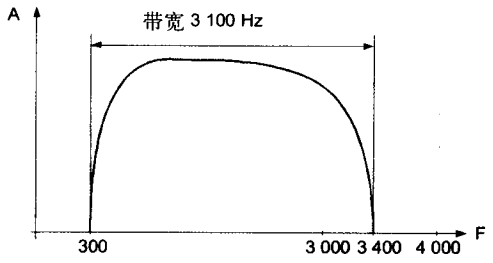


图9-2　音频信道的幅频特性

传输离散信息所用的主要调制方法如图9-3所示。图9-3a显示了源信息的比特序列，用高电平代表逻辑1，低电平代表逻辑0[⊖]。这种编码被称做*电平编码*（*potential code*），并在计算机内部不同单元间传输数据时频繁使用。

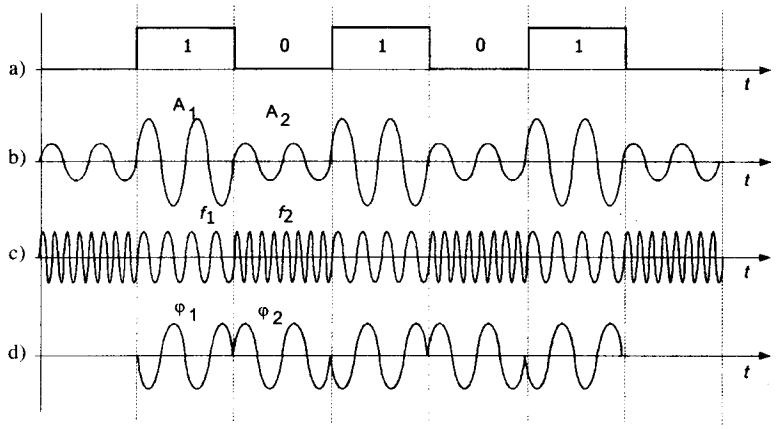


图9-3　调制的不同类型

当使用ASK方法（图9-3b）时，载波振幅的不同级别被用来指定逻辑1和逻辑0。因为它拙劣的抗噪性而很少使用。尽管如此，它却经常和PSK一起使用。

使用FSK时（图9-3c），以不同频率——f_0和f_1正弦波传输源数据中的0和1值。这种调制方法不需要在调制解调器中实现复杂的电路，通常使用在传输率为300或1 200b/s的低速调制解调器上。当只使用两种频率时，每一时钟只传输1比特，因此，这一方法称做*二元FSK*（*BFSK*）。也可以使用四倍频率用于每个时钟编码2比特。该方法称做*四级FSK*（*four-level FSK*）。另一术语，*多极FSK*（*MFSK*）也常使用。

使用PSK（图9-3d），相同频率不同相位（例如：0°和180°，或者0°，90°，180°和270°）的信号对应于0和1。第一种情况下，称做二元PSK（BPSK）；第二种情况称做积分PSK（QPSK）。

9.2.3　组合调制方式

为了增加数据率，使用组合调制方法。最常用的**积分调幅**（quadrature amplitude modulation,

⊖ 通常，电平编码使用两个不同的值表示逻辑1和逻辑0——正电平对应1，负电平对应0。

QAM）方法基于调相和调幅的组合。

图9-4展示了其中的16-QAM调制情况，这里使用了8个不同相位和4个振幅。尽管如此，只使用了32个信号组合中的16个，因为临近相位的振幅允许值不同。这样可以改善编码的抗噪性能但要降低信息率两倍。另一种改善编码可靠性却以冗余为代价的解决方式是**窗格编码**（*trellis code*）。这种编码在每4位信息中加入第5位。这一额外的比特位能在错误发生时以非常高的概率验证4信息比特集合的正确性。

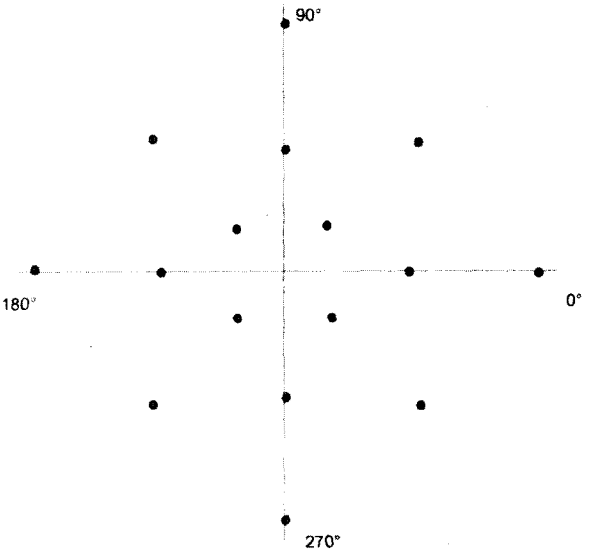


图 9-4

调制后信号的频谱依赖于调制的类型和速度（例如，依赖于源信息的适当传输率）

首先，使用电平编码时应考虑信号频谱。假设逻辑1用正电平编码，而且逻辑0用相同大小的负电平编码。为简单起见，假设传输信号由交替变化的1和0无限序列构成，如图9-3a所示。

频谱可由用于周期函数的**傅里叶公式**（Fourier formula）[⊖]直接得到。如果离散信息以比特率Nb/s传输，那么频谱由零频率的直流部分和频率为f_0，$3f_0$，$5f_0$，$7f_0$等的无限谐波函数序列组成，这里$f_0 = N/2$。频谱的频率f_0称作**基础频率**。

谐波的振幅下降得非常缓慢——相对于f_0的谐波振幅而言，因子为1/3，1/5，1/7等（图9-5a）。结果，为了确保高质量的传输，电平编码频谱需要宽带宽。除了这些以外，有必要注意实际当中，信号常常根据线路上传输的数据发生变化。例如，长序列的1和0传输会使频谱向低频率移动。在极端情况下，当传输的数据只由1（或者只由0）构成时，频谱将由零频率的谐波构成。当传输交替的1和0时，没有直流部分。因此，在二进制数据传输过程中，电平编码最终信号的频谱，其频率范围从接近0Hz的值到频率大约$7f_0$的谐波，高于$7f_0$的频率则忽略，因为它们对于最终信号几乎没有意义。对于音频信道而言，电平编码的上限是971b/s传输率，任何速率的下限都是不可接受的，因为信道开始带宽为300Hz。因此，电平编码从不在声道中使用。

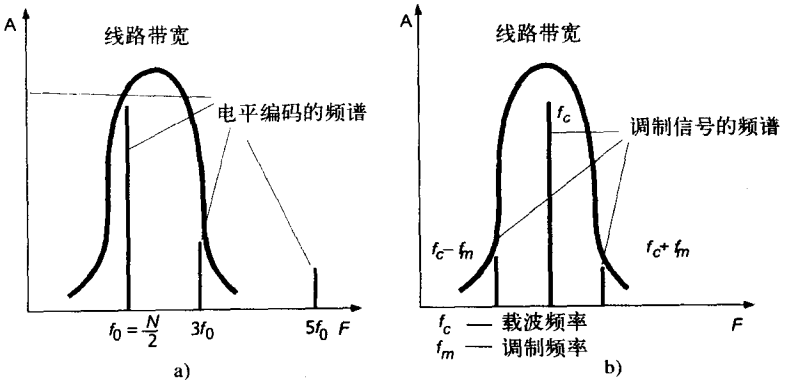


图9-5 使用电平编码和调幅的信号频谱

⊖ 参考大学数学教材的相关章节。

使用AM时，信号频谱由载波频率为f_c的正弦波，两横向谐波$f_c + f_m$和$f_c - f_m$，以及两横向谐波$f_c + 3f_m$和$f_c - 3f_m$构成。这里，f_m是正弦波信息参数变化频率，其与使用两个振幅级时的传输率一致（图9-5b）。f_m决定了所选编码方法的链路带宽。使用小的调制频率时，信号频谱带宽也小。事实上，它等于$2f_m$，如果使用$3f_m$的谐波且忽略更低能量的频率。

使用相位和频率调制，比起上面第二种情况下使用的AM方法，信号频谱会更加复杂，因为会有更多的横向谐波。尽管这样，它们对称分布在主载波频率上，且它们的振幅迅速降低。

9.3　数字化模拟信号

在本节中，我们集中解决转换问题，即以数字形式传输模拟信号。

实际上，这一问题已经于20世纪60年代解决了，当时电话网络开始传输1和0形式的声音。如果声音在传输过程中严重失真，使用模拟方式就不能保证传输质量，这是需要进行转换的主要原因。模拟信号本身不能提供失真发生的指示，或修正失真的命令。这一情况经常发生，因为信号可以有任意波形，其中包括接收端注册的波形。提高线路质量（特别是与区域载波相关时）需要很多努力和投资。因此，声音记录和传输的模拟设备被数字设备取代。这一技术使用对源模拟过程的即时脉冲调制。

9.3.1　脉冲编码调制

考虑**脉冲编码调制**（pulse code modulation, PCM）实例中的脉冲调制原则，它在数字电话中广泛使用。

脉冲调制方法基于同时使用振幅和时间的连续过程的采样（图9-6）：

- 使用预定义的周期度量连续源函数的振幅。这允许用户进行实时采样。
- 然后，用特定带宽的二进制数表示测量值，依次代表量化的函数值——振幅可能值的连续值集由这些值的离散值集替代。

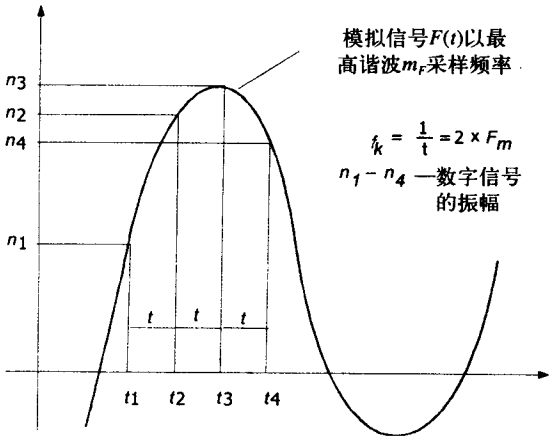


图9-6　连续过程的离散调制

实现这一功能的设备称做**模拟-数字转换器**（ADC）。之后，通过通信链路将采样值以一系列1和0的形式传送。同时，在传输初始离散信息（例如，基于B8ZS或2B1Q的编码方法）时使用相同的编码方法，这些将在本章稍后讨论。

在线路的接收端，编码转化为源比特序列，使用专门的设备，称为**数字-模拟转换器**（DAC），完成连续信号的数字振幅解制，因而重储源连续时间函数。

脉冲调制基于**奈奎斯特-科尔尼科夫信号采样原理**（Nyquist-Kotelnikov Signal Sampling Theory）。根据这一理论，如果满足采样频率是源函数频谱最高谐波频率的两倍或两倍以上，以时间采样值序列形式传输的模拟连续函数可以无损重建。

如果不能严格遵守这一条件，重建函数将和源函数明显不同。

使用数字方法记录、再生和传输信息的优点在于可以控制从介质读取或使用通信线路接收数据的可靠性。为了实现这一目的，可以采用与处理计算机数据相同的方法，例如计算校验和、重传失真帧和使用自纠错码。

9.3.2 数字化声音

对高质量的声音传输，PCM方法采用8 000Hz的声音振动振幅为采样频率。这是由于在模拟电话中，300到3 400Hz用来传输声音数据。这一范围允许以足够的质量传输人类声音的主要谐波。根据奈奎斯特-科尔尼科夫原理，对于高质量的声音传输，选择超过连续信号最高谐波至少两倍（例如，$2 \times 3\,400 = 6\,800$ Hz）的采样频率已经足够。实际选择的采样频率（8 000Hz）提供了质量预留。因此，PCM使用7位或8位编码代表每个采样。这些值对应于127级或者256级的声音信号，证明这对于高质量的声音传输已足够了。

使用PCM方法，根据代表每一采样的比特数，声音信道需要56Kb/s或64Kb/s的带宽。如果使用7位实现这一目的，那么采样率为8 000Hz，结果为：

$$8\,000 \times 7 = 56\,000\text{b/s}或\ 56\text{Kb/s}$$

相应地，对于8位，结果为：

$$8\,000 \times 8 = 64\,000\text{b/s}或64\text{Kb/s}$$

64Kb/s数字信道，也称做**元数字信道**（elementary digital channel），是标准信道。

以离散方式传输连续信号需要网络严格遵守两次连续采样之间的时间间隔为125μs的要求。这一间隔对应于8 000Hz的采样频率。这意味着网络必须在网络节点间同步地传输数据。如果接收到的采样没有同步，那么源信号就会不正确地恢复，这样就会使声音、图像或者其他用数字网络传输的多媒体信息产生失真。因此，10ms同步失真可能导致回音影响，200ms或更多时间的延迟将导致单个字符无法识别。同时，单一采样的缺失实际上对于声音再生质量没有影响，如果其余的所有信号都同步到达的话。这是由于DAC中的平滑设备基于任何物理信号的惯性属性。在这种情况下，声音振动的振幅不会马上显著地改变一个值。

DAC之后信号的质量不仅受到达输的采样同步的影响，也受这些采样量化错误的影响。在奈奎斯特-科尔尼科夫原理中，假设函数振幅可以精确测量。但是，使用二进制数的有限带宽使振幅失真。相应地，重建的连续信号也发生失真。这一现象称为量化噪声。

9.4 编码方法

9.4.1 选择编码方法

选择编码方法时，应该同时满足多个目标：

- 作为编码结果的信号应最小化频谱带宽
- 确保发送端和接收端间的同步
- 确保抗噪声
- 确保比特错误检测，如可能，进行错误纠正
- 最小化发送端能量

信号频谱（*signal spectrum*）在第8章已经讨论过了，它是编码方法的最重要特性之一。狭窄的信号频谱允许在相同链路（有相同带宽）上有更高的数据传输率。通常情况下，信号频谱依赖于编码方法和发送端时钟频率。例如，假设我们已经开发了两种编码方法，每一种方法都在一个时钟（即二进制编码）内传输1比特信息。同样，假设在第一种方法中，信号频谱带宽F等于信号的时钟频率f（如，$F = f$），第二种方法保证$F = 0.8f$。然后，使用相同带宽B，第一种方法允许的数据传输率为B b/s，而第二种方法的数据传输率等于$B/0.8 = 1.25B$ b/s。

*发送端和接收端同步*需要确保当必须从通信线路上读取新信息时接收端可以感知。传输离散信息时，时间总是被分割成相同持续时间的时钟周期，接收端在时钟周期的中段读取每一信号

(即用发送端同步它的动作)。

网络中的同步问题比紧密连接在一起的设备（例如计算机内的各种单元之间或者计算机和本地打印机间）中的数据交换复杂的多。对于较短的距离，建立在分开的时钟通信线路上的设置(图9-7）工作得非常出色。根据这一方法，只在时钟脉冲到达时从线路上读取信息。

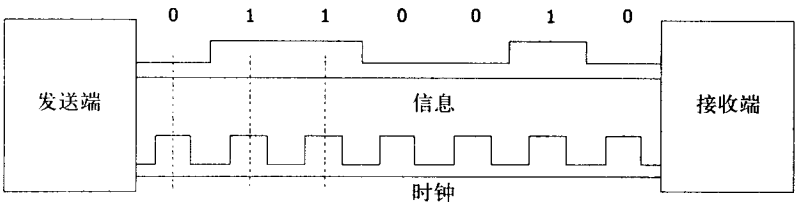


图9-7　短距离发送端和接收端的同步

在网络中，因为电缆导体属性的不一致性使用这一方法十分困难。长距离时，信号传播速度的不一致性导致了时钟脉冲的到达显著地早于或晚于对应的数据信号，致使数据位被忽略或者读取两次。在网络中摒弃时钟的另一个原因是电缆导体太昂贵。

在网络中，使用**自同步编码**（self-synchronizing code）解决这一问题。这类编码的信号包含接收端指令，说明识别出下一位（或者多位，如果编码是面向多于两个状态的信号）的时间。任何信号电平的下降沿——所谓“前沿（front)”——是作为接收端和发送端间同步的最好表征。

当使用正弦波作为载波信号时，结果信号具有自同步属性。这是由于载波频率的振幅变化允许接收端探测下一时钟的开始时刻。

失真数据的检测和纠正很难使用物理层提供的方法实施。因此，这一工作常使用更高层协议实现：数据链路层、网络层、传输层或应用层。另一方面，物理层的检测可以节约时间，因为接收端不必等到整个帧全部载入缓存，而是当检测到其中有错误位时马上丢弃。

编码方式的需求相互矛盾。因此，这一章中讨论的流行编码方式各有优缺点。

9.4.2　电平不归零码

图9-8a显示了**电平编码方法**（potential encoding method)，也称做**电平不归零**（nonreturn to zero encoding, NRZ）**编码**。后一个名字反映了当传输1的序列时，信号在时钟周期内不返回为零，这与其他编码方式不同。

NRZ编码方法有如下优点：

- 容易实现
- 良好的错误检测能力（因为两个不同电平）
- 狭窄的频谱（因为基础谐波f_0的频率低于N/2Hz，这已在9.2.3节中说明）

不幸的是，这一方法存在不可避免的缺点，主要如下：

- 缺少自同步特性。即使接收端有高精度的时钟振动，也可能在决定实例数据读取时发生错误，因为两次振动的频率从来不相同。同时，当传输长序列的1和0时，线路上的信号不会改变。因此，以高数据交换率，假设传输长序列的1和0，甚至一个很小的发送端和接收端频率不和都可能导致相当于整个时钟的错误。结果，接收端会读取出不正确的比特值。
- 如果传输长序列的1和0，那么会存在接近零的持续低频部分。因此，许多通信链路不提供在不支持此编码方法在发送端和接收端之间进行的直流连接。因此，纯NRZ格式的编码不在网络中直接使用。尽管如此，网络使用它的一些变化形式。这些变化形式消除了NRZ编码的自同步劣势和持续低电平部分带来的问题。

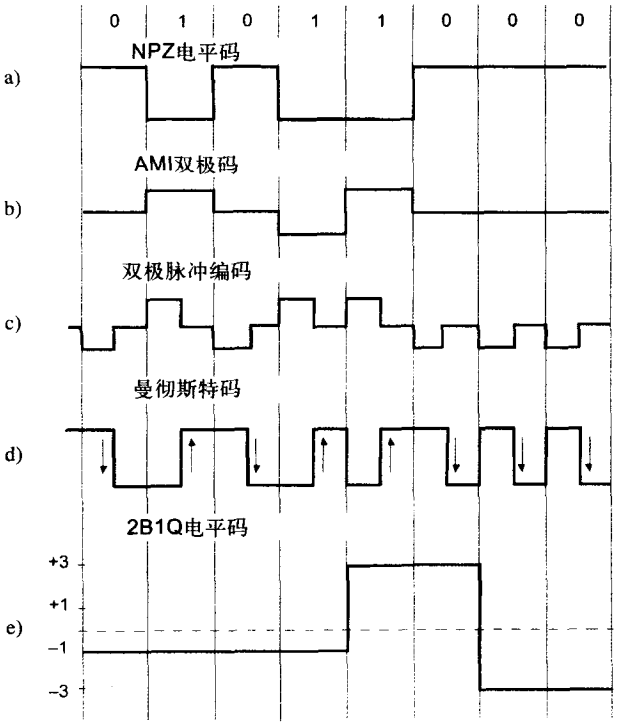


图9-8 离散数据编码的方法

9.4.3 双极标记交替反转编码

双极标记交替反转编码（bipolar alternate mark inversion, AMI）是NRZ编码的变化形式之一。这种方法（图9-8b）使用三级电平——负、零、正。零电平用于编码二进制零（即遇“0”码时为零电平），二进制电平由交替转换极性的非零脉冲编码（即遇“1”码则交替转换极性，对应极性交替变换的正、负电平）。

AMI编码一定程度上消除了直流分量问题，但缺乏NRZ编码的自同步特性。这在传输由“1”组成的长串时发生。在这种情况下，线路上的信号是一系列交替变换极性的脉冲，其与用于发送交替的“1”和“0”的NRZ码（即：无直流分量）具有相同的频谱，且主谐波频率为$N/2$ Hz。至于传输由“0”组成的长串，AMI编码并不比NRZ编码有更高的安全性，因为在这种情况下，信号退化为持续的零振幅电平。

整体而言，对于不同的比特组合，使用AMI编码会比使用NRZ编码具有更窄的信号频谱。因此，AMI编码增加了线路容量。例如，当传输交替变化的“1”和“0”时，主谐波的频率f_0为$N/4$ Hz。AMI编码也提供了错误检测能力。例如，不符合信号极性交替变换规则的脉冲都被视为错误脉冲或者从线路中缺失了正确脉冲。

AMI编码使用三级信号而不是两级。额外的信号级要求发送端功率增加大约3dB以保证位接收具有相同的可靠性。这是所有使用多信号状态编码的共同缺点。

9.4.4 “1”翻转的不归零码

这是与AMI编码相似但却只有两个信号级的编码方法。当传输“0”时，编码传输与前一时钟周期相同的电平（即不改变电平），而当传输“1”时，编码改变电平。该编码方式称做**‘1’翻转**

的不归零码（nonreturn to zero with ones inverted, NRZI）。当不使用第三个信号级时非常方便，例如，在光纤电缆中，只有两种信号状态可以稳定检测——明和暗。

使用两种方法可以改善与AMI和NRZI相类似的电平编码。第一种方式基于在源码中添加包含逻辑1的冗余位。在这种情况下，0长串被打破，编码获得了为任意传输数据自同步的特性。直流部分也被忽略了，这意味着信号频谱狭窄。尽管如此，这种方法降低了有效带宽，因为冗余1不携带任何用户信息。

另一种方法基于将原始信息混合，使得0和1出现的概率近似相同。实施这一操作的设备或单元称做**扰频器**（scrambler）。在扰频过程中，使用了一个著名的算法；因此，接收端在接收二进制数据后，把它们传输给**解扰器**（descrambler），用来恢复初始比特序列。

9.4.5　双极脉冲编码

除了电平编码外，网络也使用脉冲编码，数据要么由完整的脉冲表示，要么由脉冲的一部分，即它的前沿表示。在**双极脉冲编码**（bipolar pulse code）中，1由电平的一极表示，而0由相反的电极表示（图9-8c），是实现这一方法的最简单编码。每个脉冲持续半个时钟周期。此编码拥有杰出的自同步特性。但是，直流部分可能存在，例如，当传输由1或者0构成的长串时。除了这些，该编码的频谱比电平编码宽。例如，当传输全1或者全0时，主谐波的频率总是等于NHz，这是NRZ编码主谐波频率的两倍，AMI编码主谐波频率的四倍。因为频谱非常宽，所以双极脉冲编码很少使用。

9.4.6　曼彻斯特编码

直到最近，**曼彻斯特编码**（Manchester code）是在LAN中最常用的编码方法（图9-8d）。使用该技术的编码用在以太网和令牌环网中。

在曼彻斯特编码中，电平变化（即脉冲前沿）用于编码1和0。使用曼彻斯特编码时，每一时钟周期分为两部分。信息由每一时钟周期中部的电平变化编码。1编码为由低信号级到高信号级的摆动，而0则表示下降。在每一时钟周期的开始，如果有必要传输多个1或者0，可能传输服务信号。因为一个数据位的传输需要信号每个时钟至少变化一次，曼彻斯特编码有很好的自同步特性。曼彻斯特编码的带宽比双极脉冲编码窄。曼彻斯特编码没有直流部分，最坏情况（即传输由1或0组成的长串）下主谐波频率为NHz。最好情况（当传输交替的1和0时）下，主谐波的频率为$N/2$ Hz，相对于AMI或NRZ编码而言。平均情况下，曼彻斯特编码的带宽比双极脉冲编码窄1.5倍，主谐波频率在$3N/4$左右摆动。曼彻斯特编码与双极脉冲编码相比有另外一个优点，因为后者使用三信号级，而曼彻斯特编码只使用两级。

9.4.7　2B1Q电平码

图9-8e展示了用于数据编码的四级电平码。这是**2B1Q编码**，它的名字反映了其主要思想——每一时钟周期传输2比特（2B或者两组）、拥有四个状态（1Q或者四组）的信号。00比特对使用−2.5V电平编码，01使用−0.833V编码，11使用+0.833V编码，10使用+2.5V编码。

当使用这一编码方式时，需要使用额外的步骤来消除相同比特对的长串，因为在这种情况下，信号变成了直流部分。当比特流随机变化时，信号的频谱是NRZ编码的一半，因为以相同的传输率，时钟持续时间变为原来的两倍。因此，使用2B1Q编码，在相同线路中数据传输率可以达到AMI和NRZI编码的两倍。但是，为了实现这一编码，必须增加发送端能量以保证四个信号状态能被接收端在有噪声背景下清晰地识别出来。

改善诸如AMI、NRZI和2Q1B等电平编码，常使用冗余码和扰频。

9.4.8 冗余码

冗余码（redundant code）的原理是把源数据序列拆分成多个部分，也称做*符号*（*symbol*）然后每个源符号用另一个拥有更大比特数目的符号代替。

举例来说，逻辑编码如4B/5B码用在FDDI和高速以太网技术中，把含有4比特的源信号用5比特的符号替代。因为结果符号中含有冗余位，因此组合后的数位通常比源数据的位长。例如，在4B/5B码中，结果字符可以包含32种组合，但是原组合只有16种（表9-1）。因此，可以从结果码中选择16种组合（那些不含有大量0的组合）且把其他组合视为*非法编码*（*code violation*）。除了保证消除直流部分和确保编码的自同步特性外，冗余位使得接收端能够检测出失真位。如果接收端遇到了非法编码，那么线路上就会发生信号失真。

分割以后，4B/5B编码通过电平编码方法在线路上传输，它只对长串的0敏感。5位长的符号可以保证线路上使用任何结果的编码符号的组合都不会超过3个连续的0。

说明 4B/5B码名字中的字母B代表“二进制”（Binary），即意味着元信号只有两个状态。也有使用三个状态的编码。例如，8B/6T编码使用6个三重符号编码8个数据位，这意味着每个符号有三个状态。8B/6T编码的冗余程度较4B/5B编码高，因为有$3^6 = 729$种结果字符，可以编码$2^8 = 256$种源编码。

表9-1 4B/5B编码的源编码和结果编码之间的对应关系

源编码	结果编码	源编码	结果编码
0000	11110	1000	10010
0001	01001	1001	10011
0010	10100	1010	10110
0011	10101	1011	10111
0100	01010	1100	11010
0101	01011	1101	11011
0110	01110	1110	11100
0111	01111	1111	11101

使用编码转换表是一项简单的操作。因此，该方法不会增加网络适配器和交换机、路由器接口部件的复杂度。

为了保证特定线路的带宽，使用冗余编码的发送端必须以更高的时钟频率操作。因此，以100Mb/s传输4B/5B编码，发送端必须工作在125MHz。同时，线路上的信号频谱比不使用冗余码时宽。但是已经证明，冗余码的频谱比曼彻斯特编码的频谱窄。这一事实也证明了逻辑编码的额外状态以及发送端和接收端都以更高的时钟频率工作。

9.4.9 扰频

扰频（*scrambling*）方法由逐位计算的结果码构成，该计算基于源编码的各位和前一个时钟周期获得的结果码的各位。例如，一个扰频器可以实现下列关系：

$$B_i = A_i B_{i-3} B_{i-5} \tag{9.1}$$

这里B_i是扰频器操作在第i个时钟周期获得的结果编码二进制数；A_i是在第i个时钟周期到达扰频器输入端的源码的二进制数；而B_{i-3}和B_{i-5}是扰频器操作在前一个时钟周期内获得的结果编码的二进制数（分别为当前时钟的前3到5个周期）且使用异或（XOR）操作（例如，模2加法）。

例如，如果源序列为110110000001，扰频器将产生如下的编码（结果编码的前三位将和源编码一致，因为所需之前的数字不可用）：

$B_1 = A_1 = 1$

$B_2 = A_2 = 1$

$B_3 = A_3 = 0$

$B_4 = A_4 B_1 = 1\ 1 = 0$

$B_5 = A_5 B_2 = 1\ 1 = 0$

$B_6 = A_6 B_3 B_1 = 0\ 0\ 1 = 1$

$B_7 = A_7 B_4 B_2 = 0\ 0\ 1 = 1$

$B_8 = A_8 B_5 B_3 = 0\ 0\ 0 = 0$

$B_9 = A_9 B_6 B_4 = 0\ 1\ 0 = 1$

$B_{10} = A_{10} B_7 B_5 = 0\ 1\ 0 = 1$

$B_{11} = A_{11} B_8 B_6 = 0\ 0\ 1 = 1$

$B_{12} = A_{12} B_9 B_7 = 1\ 1\ 1 = 1$

因此，下列串作为扰频器的输出：110001101111，它不包含在源编码中六个连续0的子串。

接收到结果串后，接收端将其传给解扰器，使用相反关系恢复原序列：

$$C_i = B_i\, B_{i-3}\, B_{i-5} = (A_i\, B_{i-3}\, B_{i-5}) B_{i-3}\, B_{i-5} = A_i \tag{9.2}$$

各个扰频算法各不相同，表现在生成结果编码数位的操作数目和编码数位们之间的替换关系。例如，在ISDN网络中，在将数据由网络传输给用户时，扰频算法使用位置5和位置23的替换转换；当由用户传输给网络时，扰频变换使用位置18和23的替换。

有更简单的方法消除连续1构成的串，也归类为扰频方法。

为了改善双极AMI编码，使用两种基于无效字符的0字符人工失真方法。

图9-9展示了使用八个0替代（B8ZS）编码和高密度双极三零（HDB3）编码用于纠正AMI编码。源编码由两长串0构成：在第一种情况下，有八个0，第二种情况有五个0。

B8ZS编码只更正由八个连续0构成的串。为了达到这一目的，编码在前面三个0后面插入五个数字：*V*-1*-0-*V*-1*。这里，V代表当前时钟周期不可用的信号极（例如，不改变前一个“1”极性的信号）。1*代表正确极性的“1”信号（星号特指在源中，此刻这里的编码为0）。因此，在八个时钟周期内，接收端会检测到两次失真。由于线路上的噪声或者其他传输错误是有可能发生的。因此，接收端会认为非法编码是由八个0构成的串引起的。接收到这些串后，接收端用八个0将其代替。B8ZS编码使用这种方法构建，他的直流部分等于含有二进制数字串的0。

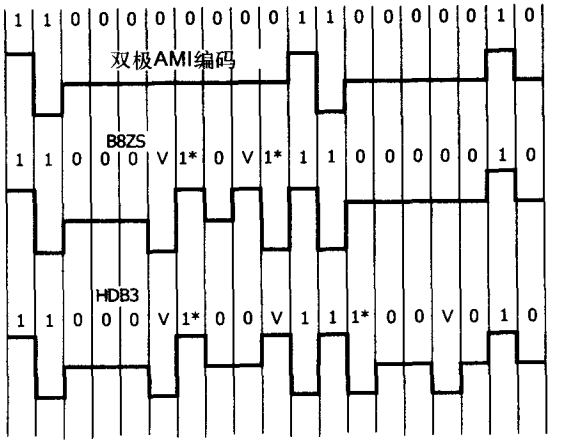


图9-9　B8ZS和HDB3编码

HDB3编码纠正源序列中四个连续0，形成HDB3编码的规则比B8ZS的复杂。每个四个0构成的组用四个信号替代，其含有一个V信号。为了抑制直流部分，V信号的极性在连续替代中变化。除了这些，还使用两种模式的四周期编码做替代。如果替换之前源编码含有奇数个1，使用000V模式；如果1的个数为偶数，使用1*00V编码替换。

改善后的电平码对于传输数据中可能遇到的传输数据的“1”、“0”序列都有很窄的带宽。图9-10展示了任何“0”“1”组合在源编码中相同可能性出现的情况下，使用不同编码方法传输数据时的频谱。建立图形时，频谱使用所有可能的初始序列集。当然，结果编码可以有不同的“1”、“0”分布。从图中可以看出，NRZ电平码有非常好的频谱，唯一的缺点——直流部分。使用逻辑编码方式获得的电平码比曼彻斯特码拥有更窄的频谱。甚至当时钟频率增加时也成立。（在图9-10中4B/5B码的频谱和B8ZS码的频谱近似相符，而不是向高频区转换，因为它的时钟频率同其他编码相比增加了四分之一。）这也解释了为何电平冗余和扰频码使用在诸如FDDI、快速以太网、千兆以太网、ISND等当前应用中，而不使用曼彻斯特或者双极脉冲编码。

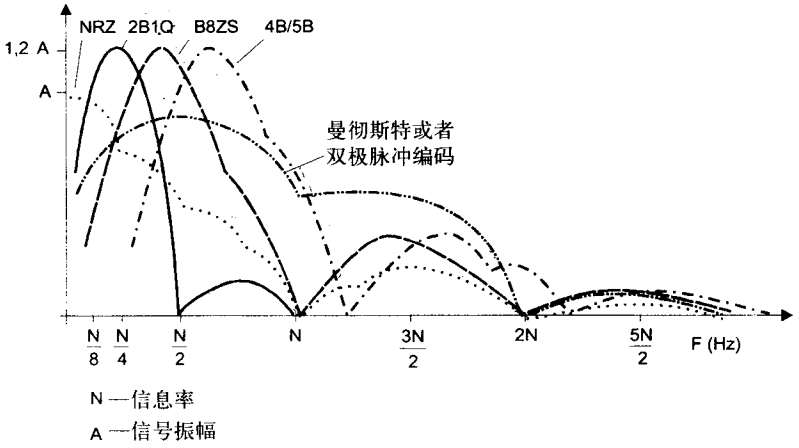


图9-10 电平和脉冲编码频谱

9.4.10 数据压缩

数据压缩（data compression）是在不损失数据内容的前提下降低原信息容量的技术。在网络中使用数据压缩用来降低传输时间。通常，由于数据压缩引起传输时间降低的好处只有在慢速链路上才能注意到。之所以发生这种情况是因为发送端需花费额外时间压缩数据，且接收端也需要花费相同的时间来解压。对于当前技术而言，传输率的阈值大约为64Kb/s。大多数网络软件和硬件工具能够完成动态数据压缩（*dynamic data compression*）而不是静态压缩，此时要传输的数据事先压缩（例如，只用著名的文档工具WinZip），然后发送到网络上。

实际中使用多种压缩算法，每一算法都应用在特定的数据类型上。一些调制解调器（智能化）提供自**适应压缩**（adaptive compression）。使用自适应压缩时，调制解调器根据传输数据选择特定的压缩算法。下面考虑几种常用的压缩算法。

十进制包装（decimal packing）意味着数据仅由数字构成，且通过降低用于编码位的比特数目可以带来可观效益。使用简单的二进制，而不是ASCII码来编码十进制数位，将把用来表示一个十进制位的比特数从7降低到4。所有对应于十进制数的ASC Ⅱ编码都明显包含001组合。如果信息帧中的所有数据只包含十进制数位，那么在帧头中放置适当的控制字符可以显著降低帧长度。

相关编码（relative encoding）是十进制包装的备选方法，当传输的数字型数据在数字间有少量偏移时使用。在这种情况下，只需传输和参考值相比的偏移值。实际上，在ADPCM数字声音编

码方法中使用这一方法，其中每个时钟周期只传输连续声音度量的差别。

符号抑制（symbol suppression）可以这样解释：通常情况下，传输数据含有大量重复的字符。例如，当传输黑白图像时，黑色表面会产生大量的0，大多数亮区域会包含大量包含1的数据。发送端扫描将被发送的字节序列，如果检测到包含三个或更多相同字节的串时，就这些串将被特殊的3字节序列替代，序列指定了字节值，重复数量，并用特殊控制字符标记序列的开始点。

变长码（codes of variable length）编码方式基于如下事实，所有的字符不是在要传输帧中都有相同的出现频率。因此，许多编码方法使用较短编码替代经常遇到的字符。较少遇到的字符则以长编码替换。这种编码方式也称做统计编码。因为符号有不同的长度，只能使用面向位的传输方式传输帧。

使用**统计编码**（statistic encoding）时，在分析位序列时选择编码来明确建立比特的特定部分和特殊符号间或者不可用位组合的对应关系。如果某种字符组合视为非法，那么就在序列中增加一位并重复分析。例如，如果选择1来表示最频繁使用的字符——E，占用一位——那么单个位的0就意味着非法。否则，用户只能编码两个字符。对于另一个频繁字符——T——可以使用01作为编码而视00为非法。对于A字符，可选用001，0001可以选来对应I。当传输符号明显有不规则频率分布时，使用变长编码最有效率。这是针对长文本串的情况。相反，当传输二进制数时效率低下，例如程序源码，因为8位编码几乎均匀分布。

基于变长的编码最常用的算法为**哈夫曼算法**（Huffmann algorithm），允许编码基于字符发生频率自动建立。哈夫曼编码的适应性变化允许在从源接收到数据时建立编码树。

许多通信设备类型，如调制解调器、桥、交换机和路由器，都支持动态压缩协议，能够降低传输信息容量的4到8倍。在这些情况下，协议保证1:4或1:8的压缩率。标准压缩协议，如V.42bis，和大量的所有权协议。实际压缩率依赖于传输数据类型。例如，图像和文本数据通常有很好的压缩率，但是对程序源编码的压缩效果不明显。

9.5　差错检测与校正

可以使用多种方法确保信息传输的可靠性。在第6章中，讨论了负责确保重传丢失或损坏数据包协议的操作原则。这些协议基于接收端在收到分组中检测损坏信息的能力。为实现这一目的使用特殊的差错检测方法。这些编码，除了在接收帧中检测差错，还具有纠错功能。后一种方法明显快于帧重传。

9.5.1　差错检测技术

差错检测方法基于在数据块中传输冗余信息。使用这一信息，可以一定程度上表述接收到的信息是否正确。在分组交换网络中，任何层的PDU都可作为这样的数据单元：帧、分组或者段。为了区别起见将它视作控制帧。

冗余服务信息称做**校验和**（checksum）或者**帧检测序列**（frame check sequence, FCS）。校验和计算作为主要信息的一项功能，没有必要只用加法。接收端根据已知算法计算帧校验和并检验是否和发送端计算的校验和相符，从而得出通过网络传输数据的正确性。除了包含源信息，还包含冗余码的编码称做码字。

有许多计算校验和的通用算法，根据它们在检测数据中错误的能力的混合程度而不同。

奇偶控制（parity control）是控制数据的最简单形式。也是功能最弱的差错控制算法，因为使用它只能检测到数据中的孤立错误。这一方法由构成控制信息所有比特位的模2加法组成。例如，如果字节为100101011，那么校验和为1。容易发现，对于含有奇数个1的信息，奇偶校验总是1，对于偶数个1，结果一定是0。计算校验和的结果是和控制信息一起传输的冗余位。如果在传输过

程中源数据（或者校验和）的任何一位发生错误，那么校验和计算结果和初始的计算结果不同，这可作为发生错误的依据。但是，诸如110101010的两次错误不会被发现，且被错误地认为是正确数据。因此，奇偶控制只用在数据的小部分（通常，为一个字节）上，它的冗余系数为1/8。这一方法很少在计算机网络中应用，因为其显著的冗余和不足的分析能力。奇偶控制有两个变量——当1的个数补足为偶数个（如同前例所示）时，使用偶校验控制，当计算结果相反，1补足数为奇数时，使用奇校验控制。

垂直和水平奇偶控制（vertical and horizontal parity control）是刚才所述方法的修改。不同之处在于将原始数据视为矩阵，其中每一行构成数据字节。控制位由每行和每列单独计算。这种方法可以检测出大多数的双重错误。但和前一方法相比，其冗余性更加明显。实际上，在网络中传输数据时从来不用此方法。

循环冗余校验（cyclic redundancy check, CRC）是当前计算机网络中最流行的差错校验方法。这一方法并不局限于网络；例如，它也在硬盘和软盘写数据时使用。这一方法认为源数据为一个多位二进制数。例如，以太网中由1 024字节构成的帧将被视为有8 192位。用已知除数作除法运算所得的余数R用作控制信息。通常，使用17位或者23位数作为除数来确保由除法得到的余数拥有16位（2字节）或者32位（4字节）长度。这包括考虑余数总比除数短一位。当接收到数据帧时，由相同除数除法运算得到的余数R被再次计算。但在这种情况下，帧中的校验和加入到帧数据中。如果余数为零，可以得出结论，接收帧没有发生错误。否则，认为帧发生了失真。

即使这种方法的分析能力比奇偶控制方法高很多，但还是很复杂。CRC方法可以检测所有独立的错误、双重错误和奇数个数位错误。这一方法的冗余性也不高。例如，对于一个大小为1 024字节的帧，控制信息长度只有4字节，相当于只有0.4%。

9.5.2 差错校正

编码技术允许接收端不仅能够检测接收数据的错误，而且还能使用称做**前向纠错**（forward error correction, FEC）的方法进行纠错。由FEC确保的编码比起只检测错误的编码需要更高的冗余性。

当使用任何冗余码时，不是所有的编码联合都是允许的。例如，奇偶控制只允许所有编码的一半。如果控制三位信息，那么下面四位编码补足奇数个1后允许使用：

000 1，001 0，010 0，011 1，100 0，101 1，110 1，111 0

这只是16种编码中的8个。

为评估差错校正所需的补充位个数，有必要了解允许编码组合的海明距离。**海明距离**（Hamming distance）是允许编码任意对比特位间的最小差别。对于奇偶控制方法，海明距离为2比特。

可以证明，如果使用海明距离为n构建冗余码，此种编码可以检测n-1种错误和纠正（n-1）/2种错误。因为奇偶控制编码的海明距离为2比特，所以只能检测到孤立错误而不能校正任何错误。

海明码（Hamming code）可以有效地检测和校正独立错误（例如，被许多正确比特分割的个体失真位）。尽管如此，如遇到长序列的失真比特（也称做差错爆发），海明码也无能为力。

差错爆发的情形会典型的发生在无线信道上。在这种情况下，使用*卷积编码方法*（*convolution coding method*）。因为为了检测大多数可能正确的编码，这一方法使用窗格图，这些编码称做*窗格编码*（*trellis code*）。这些编码不仅使用在无线信道上，而且还用在调制解调器上。

9.6 多路复用和交换

编码和差错校正方法允许用户在传输介质上创建传输链路，诸如铜电缆线。但这不足以有效连接网络用户。在链路内，需要创建独立的信道用于交换用户的信息流。对于创建用户信道，传

输网络的交换机必须支持一些多路复用和交换技术。交换技术与用于创建信道的多路复用方法密切相关；因此，这一章将涵盖这两个方面。

目前，如下方法用于多路复用用户信道：

- 频分复用（FDM）
- 波分复用（WDM）
- 时分复用（TDM）
- 码分多路访问（CDMA）

TDM同时使用电路交换和分组交换技术。诸如FDM、WDM和CDMA方法只用电路交换技术。CDMA方法只在扩频方法时使用，且在下一章中有更详尽的介绍，它专用于无线传输。

9.6.1 基于FDM和WDM的电路交换

频分复用技术（FDM technique）用于电话网络，虽然也适用于其他类型的网络。一些实例为传输网络（微波信道）或者电缆电视网络。

这一方法的主要思想是在链路的公共带宽上为每个连接分配专用的信道。

基于子带，创建**信道**（channel）。通过使用上述描述的方法之一调制的信道，并使用属于该信道带宽的载波频率传输数据。使用频率混合器多路复用实施，使用一个与信道带宽相同的窄带过滤器完成解多路复用。

下面讨论一下在电话网络中该种多路复用的特殊属性。

电话网络用户的源信号发送给频分复用交换机的输入端。交换机通过调制特定的载波频率改变每一信道所分配带宽的频率。为了防止不同信道的低频率部分相互混合，带宽分配为4KHz而不是3.1KHz，在它们之间允许有特定的900Hz保证间隔（图9-11）。两个FDM交换机间的链路同时传输用户信道信号，虽然这些信道都有自己的频率带宽。这样的链路称做多路复用链路。

输出FDM交换器为每个载波频率分配调制信号，并把它们传输给直接与用户电话机相连的输出信道。

FDM交换机可以同时实施动态和永久交换。当使用动态交换时，用户通过发送被呼叫用户的电话号码到网络中来初始化与另一个用户的连接。交换机动态地将可用多路复用信道中的一条分配给这个用户。当使用永久交换时，通过适当地调整交换机，把4KHz的带宽分配给用户一段时间。

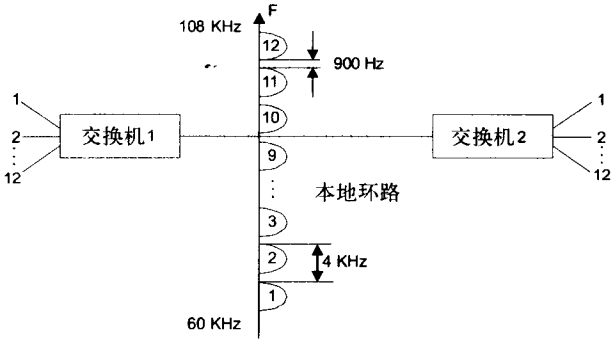


图9-11　基于FDM的交换

在其他类型的网络中，FDM交换的规则没有变化。只将带宽边界分配给特定的本地环路，同时许多在高速信道中的低速信道也发生了变化。

波分复用方法（WDM method）在电磁频谱领域使用与FDM相同的规则。这里的信息信号既不是电流也不是电磁波；相反，是光。为了在光纤电缆中组织WDM信道，使用来自三个透明窗口的电波，其对应于红外线波段波长范围为850nm到1 565nm，也即频率范围从196到350THz。

在主干链路中，多个光谱信道被多路复用——16，32，40，80或者160（从16信道开始）。这一多路复用技术称做**密集波分复用**（Dense WDM, DWDM）。在这些光谱信道中，数据用离散方式或者模拟方式编码。WDM和DWDM是模拟频率多路复用思路不同形式的实现。WDM或DWDM网络与FDM网络间的差别是信息最大传输率。FDM网络通常可以在主干网上保证600路会话的同时传输，这对应于36Mb/s的总传输率。相比而言，在数字信道中，传输率基于每个信道64Kb/s传

输来计算，DWDM网络通常确保每秒数百吉比特甚至钛比特的吞吐量。

DWDM技术将在第11章中详细说明。

9.6.2 基于TDM的电路交换

基于FDM的电路交换技术用来传输代表声音的模拟信号。当将声音表示为数字形式时，发展出了面向传输离散属性信息的多路复用技术。

这一技术即*TDM*。TDM规则由为每个连接分配信道的一段时间构成。使用两种TDM类型：异步和同步。你已经熟悉**异步TDM模式**（asynchronous TDM mode），因为其应用于分组交换网络。每个分组占用信道的一段时间以便在信道节点间传输。不同信息流间没有同步机制，每个用户在需要传输信息时尝试使用信道。

这节，我们考虑**同步TDM模式**（synchronous TDM mode）⊖，当所有的信息流同步访问信道。因此，每一信息流周期性在其配置内占有信道一段固定时间，也称做**时间槽**（time slot）。

图9-12阐明在声音传输实例中基于TDM方法的电路交换规则。

TDM网络设备——多路复用器、交换机和解多路复用器——以时间共享模式进行操作，在操作周期内为所有信道轮流服务。TDM设备的操作周期为125usec，这等于在数字信道上测量连续声音的时间段。这意味着多路复用器或交换机有时间为任意信道服务，且可以使用网络传输更深层连续度量。为每个连接分配一定量的设备操作周期，也即前面提到的时间槽。时间槽的长度依赖于TDM多路复用器或交换机服务的信道数量。

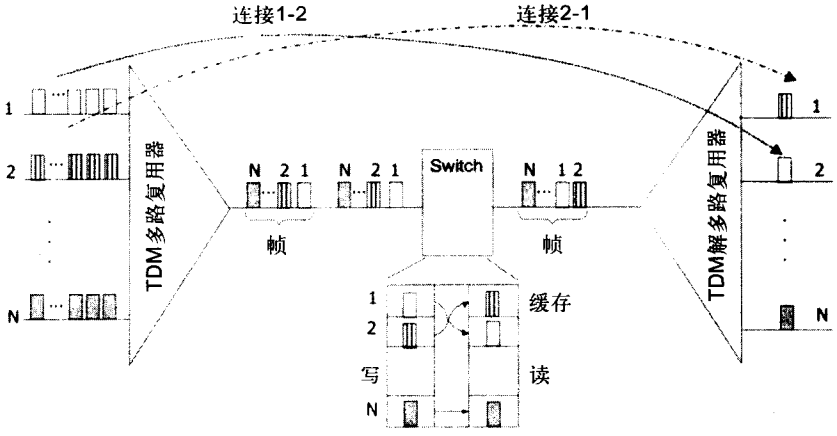


图9-12 基于TDM的电路交换

多路复用器（*the multiplexer*）通过N个输入信道接收信息，每个信道以64Kb/s（例如，每125usec一个字节）速率传输数据。在每个周期内，多路复用器实施以下操作：

- 从每个信道上接收下一字节数据
- 用所接收字节创建一个帧
- 以比特率为$N \times 64$Kb/s将多路复用帧传输至输出信道

多路复用帧中的字节顺序与接收字节输入信道上的数目相同。多路复用器服务的信道数量取决于它的操作速度。例如，T1多路复用器，是第一个基于TDM技术操作的工业化多路复用器，支持24路输入信道并输出*T1标准基础帧*（*T1 standard fundamental frame*），并以1544Mb/s比特率传输。

解多路复用器（*demultiplexer*）执行相反的工作。它解析构成帧的字节并将它们分配给输出信道，并且考虑到基础帧中的字节顺序数与输出信道的数对应。

⊖ 值得一提的是，当TDM缩写没有指出操作模式时，总指同步TDM模式。

交换机（*the switch*）从高速信道上接收来自多路复用器的帧，并将帧中的每一字节按打包入帧的顺序写入缓存的独立单元中。对于交换，按专用信道连接的用户数字来提取缓存中的字节，而不是按它们目的地的顺序。例如，网络左手边的第一个用户（图9-12）连接到网络右手边的第二个用户，所以写入缓存第一个单元的字节将是第二个提取的字节。通过把帧中的字节以需要的顺序混合，交换机保证了用户间的连接确立。

分配的时间槽数是整个连接过程中输入信道和输出时间槽之间的连接配置，即使在传输流量是突发性的和不总是需要分配时间槽数的条件下也是这样。这就意味着TDM网络中的连接总是拥有已知和固定的带宽，也即64Kb/s的倍数。

TDM设备的操作与分组交换网络的操作相似，因为每个字节都可以看做是一个元分组。尽管如此，与计算机网络中的分组相比，TDM网络中的分组没有独立的地址。地址的角色由基础帧中的字节顺序数或者多路复用器和交换机中的时间槽分配数来代表。使用TDM技术的网络需要所有设备同步操作，因为这一技术的第二个名字为**同步传输模式**（Synchronous Transfer Mode, STM）。

违背同步会破坏用户间的设备交换，因为地址信息丢失。因此，在TDM设备中不同信道间动态重分配时间槽是可能的，即使在多路复用器操作的特定周期内，特定信道的时间槽是冗余的，因为这一信道的输入端没有数据传输。一个例子就是电话用户保持沉默。

有一个TDM技术的修改，也称为**统计TDM**（statistical TDM, STDM）。特别开发这一技术是用来在万一有些信道的时间槽暂时可用的情况下，增加其他信道的带宽。为了实现这一任务，每一数据字节以简短的地址域补足（例如，4位或5位），这将允许16路或32路信道多路复用。STDM是拥有简单寻址和较窄应用领域的分组交换技术。STDM技术的应用不是非常广泛，主要用在连接终端到主机的非标准设备上。**异步传输模式**（asynchronous transfer mode, ATM）技术是静态多路复用思想的进一步发展。ATM为分组交换技术。

TDM网络可支持动态交换模式，永久交换模式，或两者皆有。例如，基于TDM技术的数字电话网络主要是动态交换。然而，这些网络还通过为用户提供租赁服务而支持永久交换。

9.6.3　信道运行的双工方式

双工方式是信道运行的最常用和有效的方式。确保双工模式的最简单变化形式是在电缆中使用两根独立的物理链路（例如，两对电线或者两条光缆），每一根都以单工模式运行（也就是，以一个方向传输数据）。在快速以太网和ATM中，这一思想视为实现许多网络技术的基础。

有时这一简单的解决方法是不可行或者低效的。更常见的是，只使用一个物理链路用于双工数据交换，第二条链路的建立需要高昂的代价。例如，使用调制解调器通过电话网络交换数据时，用户只有一条物理链路连接到自动交换机——双线。再购买另一条在经济上是低效的。在这种情况下，使用FMD或TDM技术将原信道分割成两个逻辑信道实现双工模式。

当使用FDM建立双工信道时，频带被分割成两部分。分割可以是对称的也可是非对称的。在后一种情况下，每个方向的信息传输率是不同的。实现这一方法的最流行的实例就是用宽带因特网访问的ADSL技术。当FDM确保双工模式工作时，也称为**频分双工**（frequency division duplex, FDD）。

当使用数字编码时，双线线路上的双工模式使用TDM技术编码。时间槽的一部分用于在一个方向传输数据，而其余部分则用于另一方向的数据传输。通常，相反方向的时间槽是交替的；因此，此种方法有时也称为“乒乓”传输。TDM双工模式也称为**时分双工**（time division duplex, TDD）。

在单股光纤的光纤电缆中，使用DWDM技术实现双工操作模式。一个方向的数据传输用某一波长的光束实施，相反方向的数据传输则用另一波长的光束完成。事实上，在单一透明窗口中的

组织两条独立光谱信道这一特殊问题的解决导致了WDM技术的发展。后来演变为DWDM。

强劲的数字信号处理器（*digital signal processor, DSP*）出现，能够在实时模式下实现复杂的信号处理算法，如果可能实现双工模式的另一种变化形式。在这种情况下，两个发送端在相反方向上同时工作，在信道上创建额外的信号。因为每个发送端知道它的信号的频谱，所以它从总信号中减去并接收从另一个发送端发出的信号。

小结

- 为了表示离散信息，使用了两种类型的信号：方形脉冲和正弦波。在第一种情况下，表示方法称做编码；在第二种情况下，称做调制。
- 传输离散信息时，用振幅、频率和正弦信号相位的变化对0和1进行编码。
- 为了提高数据率，使用组合调制方法。最常用的是积分振幅调制方法（*quadrature amplitude modulation method*）。这些方法基于相位和振幅的组合调制。
- 当选择编码方式时，有必要同时实现多个目的：
- 最小化结果信号的频谱的可能宽带
  - 确保发送端和接收端间的同步
  - 确保抗噪性
  - 检错，如可能，纠错
  - 最小化传输能量
- 信号频谱是编码方法的最重要特性之一。窄信号频谱允许实现介质固定带宽的更高的数据传输率。
- 编码必须有自同步属性，这意味着它的信号必须包含指示符，接收端可根据这一指示符决定在什么时候有必要实施下一个比特的识别。
- 使用离散编码时，二进制信息用不同的恒定电平级或脉冲极性表示。
- 不归零（NRZ）码是最简单的电平码。尽管如此，它不提供自同步特性。
- 为了改善NRZ电平码，使用特殊方法实现同步。这些方法基于：
  - 在源数据中引入冗余位
  - 源数据扰频
- 海明码和卷积码不仅允许检错，还可用于重复错误纠正。这些编码常用在前向纠错工具中。
- 为了提高网络中的有效数据率，使用基于不同算法的动态数据压缩。压缩率取决于数据类型和所使用的算法，压缩率的范围由1：2到1：8。
- 为了在传输链路中建立多个信道，使用多种多路复用方法：频分复用，时分复用，波分复用和码分多路访问。分组交换技术可以和TDM兼容，但电路交换技术可以使用任何多路复用技术。

复习题

1. NRZ编码的优、缺点分别是什么？
2. 什么类型的信息传输使用ASK？
3. 为什么ASK调制方式不用在宽带信道中？
4. 使用QAM方法时，正弦变化的参数是什么？
5. 用七个状态编码一个字母时，传输要使用多少比特位？
6. 解释64Kb/s带宽用来选作数字电话网络基础信道的原因。
7. 使用什么方法改善B8ZS编码的自同步属性？

8. 逻辑编码和物理编码间的差异是什么？
9. 什么规则视为差错检测和校正的基础？
10. 列出最适合文本信息的压缩方法。为什么这些方法对于二进制数据压缩是低效的？
11. 海明距离是什么？
12. 对于奇偶控制方法，海明距离取值为多少？
13. 在以太网中使用频分复用是否可行？
14. 在分组交换网络中使用何种TDM模式？
15. 是否可结合多种多路复用方法？如果可以，给出适当的实例。
16. FDM和WDM方法的共同点是什么？
17. 如果信道传输双方同时使用相同的频率范围，那么双工模式将基于何种技术？

练习题

1. 找出NRZ信号频谱的最初两个谐波，当传输序列为110011001100……如果传输频率为100MHz。
2. 使用3B/4B码时，你会选择哪16个编码来传输用户信息？
3. 建议一种海明距离为3比特的冗余码。
4. 如果使用如下参数：载波频率为2.1005GHz，两振幅值的ASK调制，时钟频率为5MHz，那么在带宽为2.1GHz到2.102GHz的信道上传输数据是否可靠？
5. 如果传输消息为：BDDACAAFOOOAOOOO。建议字符A、B、C、D和O采用变长码。对于如下方法是否可以实现压缩：
   - 传统ASC Ⅱ码？
   - 只考虑前面列出字符的固定长度编码？
6. 如果发送端时钟频率翻倍，NRZ信号频带可以改善多少倍？

第10章 无线传输

10.1 引言

在使用导向介质传输几年后，无线传输成为通信的流行方式。到19世纪90年代，发明出第一个在传输电报领域使用无线电信号的设备。到20世纪20年代，人们开始使用无线电进行声音传输。

如今，已经有许多不仅仅限于广播，如电视和收音机的无线通信系统。无线系统广泛地用于传送离散信息。为了建立长距离的通信链路，可使用无线电中继和卫星系统。无线访问系统也可用于访问通信载波网络和无线局域网。

无线介质主要使用频谱范围，且以高噪声为特征。噪声产生于外部放射源以及从墙体和其他障碍物多次反射的信号，因此无线通信系统需要实施多种抗噪方法。这些方法包括已经讨论的前向纠错码，以及带有信息传输应答的协议。专门为无线系统开发的扩频技术是一种有效的抗噪工具。

这一章中，我们提供关于建立点对点、点对多点和多点对多点通信信道所使用的无线系统的元素，操作原理和编码方法。

10.2 无线介质

10.2.1 无线通信的优点

不使用电线即可传输信息成为可能，把用户从局限的特定地点解放出来，一直拥有吸引人的前景。只要技术上能够满足崭新的无线服务获得成功所需的两个组成部分——方便使用和低廉的成本，那么就可以保证它的成功。



图10-1 第一个汽车电话
[作者——Anders Sunesson纪许及表]

移动电话是这一点的新证实。第一部移动电话由*Lars Magnus Ericsson*在1910年发明。发明的电话用于在汽车上使用，但只能在路途中实现无线。在这些条件下，不可能使用设备。若要打电话，就需要停车。然后，使用一组长杠将电话连接到路边的电话线（图10-1）。显然，这既不方便使用又限制了移动性，这产电话不可能获得商业成功。

无线通信并不只意味着移动无线通信。也有固定无线通信（fixed wireless communication）。在这种通信中，通信节点始终位于小块地域的边界内，例如，在特定的建筑、地形或区域内。

如果因为某些原因使用电缆系统不可实行或低效，那么可由固定无线通信替代导向介质通信。这些原因各不相同。人口稀疏区或者难以到达的地区，诸如沼泽带、巴西丛林、沙漠、北极和南极区不得不为它们的电缆系统等待相当长的时间。有历史纪念意义的建筑和会被电缆危及其安全的墙体是另一实例。另一常见的情景是有必要为用户提供连接，但他的家已经连接到现有已分配载波的节点上。最后，有时需要建立一些临时通信。例如，在没有铺设电缆的建筑物内招开会议，但要确保为所有与会者提供高质量的通信服务时，可使用无线通信。

将无线通信应用于数据传输已有很长时间了。直到最近，在计算机网络中的大多数无线通信应用都还局限于固定类型。网络用户甚至网络设计者不会总意识到，在网络路径的某些部分，信

息是不能够通过线路传输的，而是以电磁波的形式通过地球大气和太空传播。这种情况发生在当计算机网络从通信技术上租赁信道，且这种信道的专业链路为卫星或陆地微波链路时。

在20世纪90年代中期，**移动网络技术**（mobile network technology）达到了所需的成熟度。随着在1997年采用IEEE802.11标准，构建移动以太网成为了可能，这确保了无关地域、设备的生产商或销售商之间的通信。此时，这一网络技术起到比移动电话网络更加现代的角色。尽管如此，大多数分析师预测，在不久的将来这一领域会有迅速的发展。

通常，无线网络和无线电信号（*radio signals*）联系起来，虽然这不总是正确的。无线通信使用非常广泛的电磁频谱。其范围覆盖从低频无线电波（几Hz）到可见光（大约8×10^{14}Hz）。

10.2.2　无线链路

依据相当简单的方法建立*无线链路*（*wireless link*）（图10-2）。

每个节点都装有天线，它的作用相当于电磁波的发送端和接收端。电磁波在大气中或真空中的传播速度为3×10^{8}m/sec。

电磁波可以在任意方向（全向的）上或在一个特定的扇区（有向的）内传播。传播类型取决于天线的种类。图10-2显示了**抛物线天线**（parabolic antenna），它是有向的。

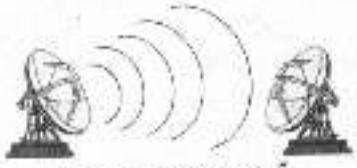


图10-2　无线传输链路

全向天线（isotropic antenna）也十分常见。它是一个具有四分之一发射波长的垂直导体。这些天线是全向的（*omnidirectional*），其广泛使用在汽车和便携式设备上。电磁波的全向传播也可以由多个有向天线实现。

在全向传播中，电磁波覆盖在由信号衰减决定的某一半径所限制的整个空间内。这一空间成为共享介质。介质共享引起与在LAN中一样的问题。但是在无线通信中，这一情况变得更糟，因为周围的空间都开放给公众访问。这与电缆不同，后者属于特定组织。

无线介质相对于导向网络而言，也称做非导向介质。导向介质中导体（铜线或光缆）严格定义了信号传播的方向。传输介质共享也会产生与局域网相同的问题。

对于使用无线信道传输离散信息，需要使用所传输的比特流对发送端的电磁振荡进行调制。这一功能由连接计算机、计算机网络的交换机或路由器和无线的DCE设备实现。

10.2.3　电磁频谱

无线通信链路的特性——节点间距离、覆盖范围和信息传输率等——许多方面都取决于所使用的电磁频谱频率。频率f和波长λ通过以下关系关联：

$$c=f\times\lambda \tag{10.1}$$

图10-3展示了电磁频谱的频率波段。当考虑电磁频谱的频率波段时，可以将无线通信系统分为以下四组：

波段范围从20GHz到300GHz称为**无线电波段**（radio band）。ITU已经将该波段分为多个子范围（图10-3中较低一行箭头），范围从极低频率（ELF）到极高频率（*extra high frequency, EHF*）。我们熟悉的无线电台，与其关联的收音机工作在20KHz到300MHz范围内。对于这些范围，使用一个广泛使用的专用名字，尽管它并没有被定义为标准——*广播无线电*（*broadcast radio*）。这一组包括低速系统，使用调幅（AM）和调频（FM），数据传输率在数十到数百千比特每秒之间。这些设备的实例为以2 400Kb/s、9 600Kb/s或19 200Kb/s速率将两个相同LAN连接在一起的无线电调制解调器。

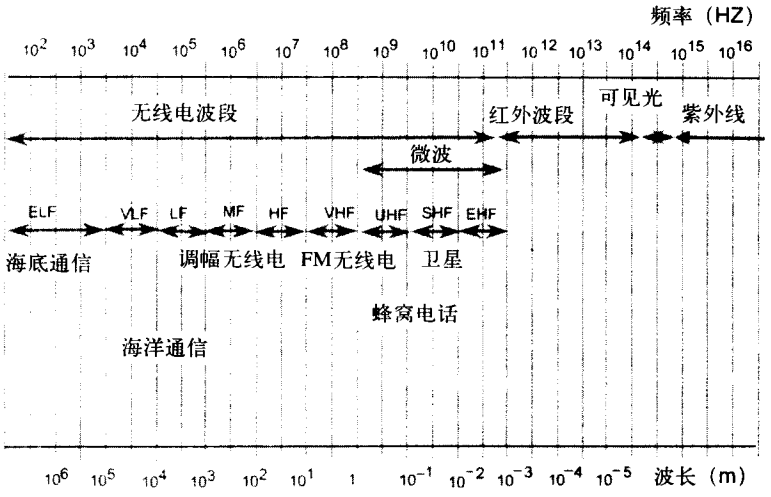


图10-3 电磁频谱的频率波段

从300MHz到3 000GHz范围也有一个非标准的名字——**微波波段**（microwave）。它是系统中最宽的一类，包括无线电中继线路、卫星信道、无线LAN和固定无线访问网络。它们也称为*无线本地环路*（*wireless local loops, WLL*）。

微波波段的上层是**红外波段**（infrared band）。微波和红外波段也在无线信息传输中广泛使用。因为这类发射不能穿透墙壁，所以这种类型的系统用于在单个房间内建立小型的LAN段。

可见光波段。在过去几年内，发现了可见光使用激光来传输信息的应用。这些系统可以作为组织短距离访问的高速点对点无线网络信道的另一选择。

说明 可见光可能是无线通信的第一介质，因为它在古代文明中（例如希腊）用于山顶上的一系列观察者之间的信号接力传输。

10.2.4 电磁波的传播

与射频相关的电磁波传播有很多模式：

载波频率越高，则可能的数据传输率越高。

频率越高，信号穿透墙壁或者其他障碍物的特性越差。AM波段的低频无线电波可以轻易地穿透住宅，因此允许家庭用户通过内置天线接收。TV信号有稍高些的频率，通常需要外置天线。最后，红外线辐射和可见光不能穿透墙壁，因此局限于使用视线传输。

频率越高，信号能量的降低将随着到发送端距离的增大而增大。当电磁波在开放空间（没有反射）传播时，信号能量衰减正比于到源信号距离平方和信号频率平方的乘积。

上限为2MHz的低频信号，沿着地球表面传播。基于这一原因，AM无线电信号可以传播数百公里。

2MHz到30MHz的信号频率可以被地球的电离层反射。因此，它们可以传播更远的距离——数千公里，只要发送端的能量足够。

频率高于30MHz的信号只能沿直线传播。这意味着它们是视线信号。超过4GHz的频率，问题也随之诞生。例如，信号可能会被水吸收，这意味着不仅雨，雾也可以明显降低微波系统的传输质量。因此，以下雨闻名的西雅图，经常实施激光数据传输测试。

快速的数据传输是首要需求；因此，所有操作在高频波段上的无线通信系统都从800MHz开始，虽然由低频波段提供的信号拥有多个优点，如可以沿地球表面传输，并从电离层反射。

为了成功地使用微波系统，也有必要考虑由以视线模式传播信号和传播过程中所遇到的障碍物所带来的问题。

图10-4说明了信号遇到障碍物后，采用以下三种机制传播：反射、衍射和散射。

当信号遇到障碍物，其波长远部分透明的，且大小明显高于信号波长，那么信号的部分能量从障碍物反射（*reflected*）。微波波段的电波波长为几厘米。因此，当信号在城市中传播时，它们部分地被建筑的墙壁反射。

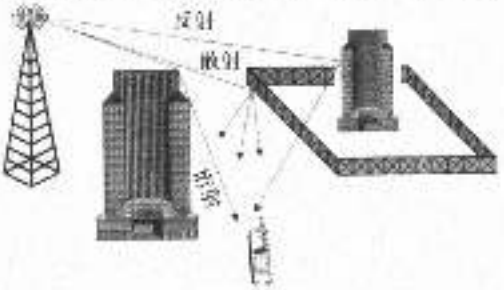


图10-4　电磁波的传播

如果信号遇到不可穿透的障碍物（例如一片金属）且其大小明显高于信号的波长，那么发生衍射（*diffraction*）。在这种情况下，信号以如下方式传播，可能在视线不可见的地方接收到信号。

最后，当信号遇到障碍物的大小与波长相当时，它开始散射（*scattered*）且以不同的角度传播。

基于这一机制，在城市中使用无线通信是十分常见的，接收端可以获得同一个信号的多个副本。这一作用称为多路传播（*multipath propagation*）。多路传播的结果往往是负面的，因为一个副本可能相位差错，抑制了主信号。通常沿不同路径的传播时间也不同，也会发生称做码间干扰（*intersymbol interference*）的现象。与这种情况关联的结果是，使编码相邻数据位的信号同时到达接收端。由多路传播引起的失真而导致的信号衰减，也称做多路衰减（*multipath fading*）。在城镇中，多路衰减导致信号的降低是距离的三次方甚至四次方而不是平方。

所有这些信号失真与外部电磁噪声相混合，在城市中数量众多甚至惊人。值得一提的是，微波炉也工作在2.4GHz波段。

> **说明**　由于不使用无线路获得的移动性需付出代价。在这种情形下，代价即无线通信链路传送中的高噪声，这是所有无线通信的通病。在希望的高频波段中，这一问题变得尤为严重。与导向通信链路中的比特误概率（10^{-9}至10^{-10}）相比，无线通信链路甚至达到10^{-3}。

无线链路上的高噪声问题可以使用多种方法解决。在宽频范围内分配信号能量的专用编码方法扮演了一个重要角色。除了这些，在高塔上放置发送端（如有可能，放置接收端）来避免多级反射也是实用方法。

使用面向连接的协议来确保协议栈中数据链路层的帧传输是另一种常用方法，其允许快速纠错，因为传输协议，例如TCP，以较大的超时值工作。

10.2.5　许可

电磁波可以在任意方向传播，可以跨越长距离传播，并且能够穿透诸如墙壁的障碍物。因此，共享电磁频谱波段的问题就变得十分紧急，而且需要集中控制。根据ITU建议，每个国家都有专门的权力授予通信运营商许可证（license），允许他们根据选用的技术使用特定波段的频谱。许可证保证在特定的地域内，拥有获授使用分配频率波段的最高权限。

授予许可时，政府权威机构使用不同的策略，最常用的策略为：

比较竞标（comparative bidding）。参与比较竞标的通信运营商要出详细的提议。在这些文档中，描述了计划服务、实现这些服务所用的技术、潜在客户价位等。委员会考虑所有的提议，并选择最能够满足社会需要和期望的公司。这样胜出者的标准的复杂性经常导致决策过程有明显证据，甚至导致官员腐败。基于这一原因，一些国家，包括美国，废除了这一方法。尽管如此，一

些国家仍在使用这种方法，更多的是在做出与关键服务相关联的决定时使用，如3G系统的采用。

抽奖法（lottery）。抽奖是最快速且公正的方法。但是，它却不能带来最好的结果。尤其是当假设运营商参与到抽奖中时（例如，那些不打算自己提供运营服务而是要出售许可证的公司）。

拍卖（auction）。如今拍卖相当流行，它能获得不公平的竞争者，且可以为国家预算带来可观的收益。这种拍卖第一次于1989年在新西兰举行。因为3G时代的临近，许多国家都为此类拍卖增加了财政预算。

有三个频率波段——900MHz、2.4GHz和5GHz——ITU建议其可不经许可[○]供国际使用。这些波段将用在使用无线通信的多用途产品上（例如，汽车门禁系统）。除了这些以外，一些科学或者医疗设备也在这一范围内工作。这一范围用引用名字命名——**工业、科学和医学**（industrial, scientific, medical, ISM）。900MHz波段最常见，因此也十分拥挤。这是可以理解的，因为特定设备的频率越高，越难保证它的低成本。高频设备往往非常昂贵。如今，正在积极发展2.4GHz波段。例如，它使用在如IEEE802.11和蓝牙等新技术上。5GHz波段的掌控刚刚开始。然而，这一波段因为可以保证高速数据传输而非常吸引人。

对于共享使用这些波段的监制要求是限制传输信号的最大能量，即不能超过1W。这一条件限制了设备使用的范围，防止对城市中其他区域的使用相同波段的用户构成干扰。

也有专用的编码方法来降低使用ISM波段的设备之间的干扰。这些方法将在下面章节中详细介绍。

10.3 无线系统

10.3.1 点对点系统

点对点有线信道的典型设计也通用于无线通信。许多个用于不同目的和使用不同频率波段的链路可以使用点对点来实现。

在传输网络中，这一设计常用于创建**无线电中继线路**（radio relay lines）。这种线路大多由多个建有抛物线天线或有向天线的塔构成（图10-5）。这种线路上的链路工作在数吉赫兹频率的微波上。有向天线以窄束聚集能量，这可以使信息传输相当远的距离（通常达到50km）。高塔保证了天线间的直接视线。

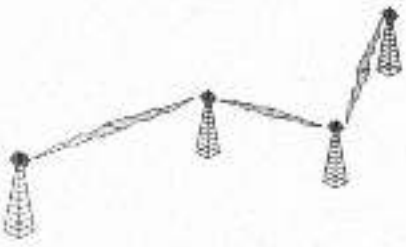


图10-5 无线电中继信道

信道带宽可能相当高。通常，从数兆比特每秒到数百兆比特每秒。这些信道可能同时代表去干和访问链路（后一情况中，通常仅由一条链路组成）。通信运营商经常在由于自然条件或经济低效而不能安装光纤电缆的条件下，使用此类信道。

相同的通信原理可用来在城市中连接两幢建筑物。因为总是需要高传输率（例如，需要连接LAN段到公司的主LAN中），在这种情况下，可以使用工作在AM频率波段的无线电调制解调器。也可使用激光连接两幢建筑物。这确保了高信息传输率——达到155Mb/s——假设大气条件很好的情况下。

另一个无线点对点信道的实例如图10-6所示。在这一实例中，它用于连接两台计算机。这些信道组成了LAN最简单的部分；因此，主要是距离和能量的不同。

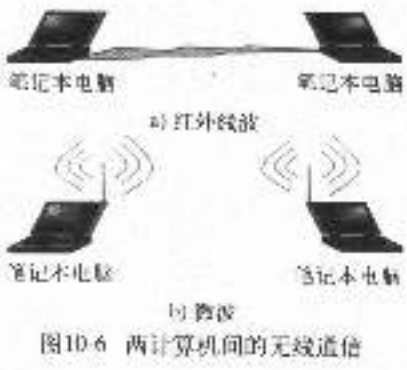


图10-6 两计算机间的无线通信

○ 与使用900MHz和5GHz频率波段相关的问题是，它们不允许是对所有国家免费使用。

对于同一房间内的通信，可以使用红外波段（图10-6a）或者微波波段（图10-6b）。大多数的当代笔记本电脑都配有内置红外线端口，允许自动建立这些连接。只需两台笔记本电脑的红外线端口在视线直线范围内或者沿着直线的反射线就可以建立连接。

室内微波通信的工作范围是数十到数百米，不可预测更远的距离，因为在室内传播的微波信号受到反射、衍射和散射的影响。这还不考虑穿透墙壁和天花板的环境噪声。

10.3.2　点对多点系统

当多用户终端连接到基站（base station）时，无线信道的点对多点设计以组织访问为特色。

无线点对多点信道同时用于固定访问和移动访问。

图10-7显示了使用微波信道的固定访问类型。通信微波使用高塔（例如电视塔）来确保装在客户建筑屋顶的天线直线可见。实际上，这种类型可能是点对点信道的集合——对应于需要连接到基站的建筑数目。尽管如此，这一方法非常浪费，因为对于新用户需要在塔上安装新的天线。基于这一原因，通常天线覆盖一个扇区（例如，45度）。在这种情况下，微波可以通过安装多个相互之间有一定距离的天线保证全向（360度）通信。

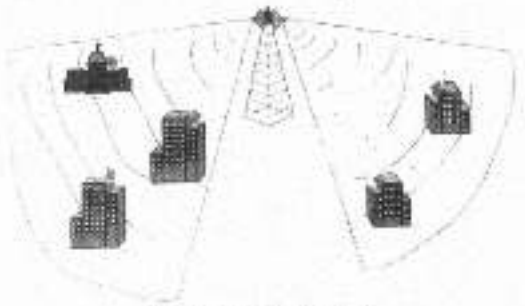


图10-7　固定无线访问

访问信道的用户只能与基站交换信息，这些基站能够作为单个用户间通信的中间节点。

基站通常与网络的导引性部分相连接，确保与其他基站用户或导引性网络用户间的通信。因此，基站常称做访问点（access point）。通常，访问点不仅包括DCE设备，该设备用于创建信道（也即物理层设备），也包括电话和分组交换机。这允许它可以像网络交换机一样工作并提供访问。

目前的大多数移动访问系统都采用蜂巢原理。这里每个**蜂巢**（honeycomb，**蜂房**（cell））代表由单独一个基站（*base station*）服务的地区。刚开始时并不使用这种思想，第一个蜂窝式电话并不基于此。早期的手机访问能够覆盖大片区域的单个基站。小蜂房思想首先在1945年提出。尽管这样，还是过了很长时间后才出现了第一个商用蜂窝式手机网络。第一个试用版诞生于20世纪60年代，到20世纪80年代早期才开始广泛应用。

把整个覆盖区域划分成小单元（蜂房）的原理叫频率重用思想完善。图10-8展示了只使用三个频率构成的蜂窝式网络组织结构。这里任何两个相邻的蜂房对之间都不使用相同的频率。频率重用允许运营商节俭地使用许可的频率范围。同时，用户和相邻蜂房的基站不会遇到相关信号的干扰。当然，基站必须控制发出信号的能量以避免使用相同频率的非相邻蜂房间的噪声干扰。



图10-8　蜂窝网络中的频率重用

当蜂房为六边形时，重用频率的数量是任意的。换句话说，并不局限于图10-8所示的三个。重用频率的数量（N）可使用3、4、7、9、12、13等。

给定使用相同频率的两个蜂房中心之间的最小距离，N可以根据以下公式进行选择：

$$N = \frac{D^2}{3R^2} \tag{10.2}$$

这里，R是蜂房半径，D为重用距离。

小的蜂房保证用户终端设备的小型和低功耗。这一情况和通用技术的进步允许创建紧凑的蜂

离式电话。

目前，移动计算机网络没有移动电话网络普及，尽管如此，它们都基于建立无线信道的相同原则。

终端设备从一个蜂房到另一个的迁移是移动信道的问题。这一过程称为传递（handoff），它并不存在于固定无线访问中。尽管如此，这一功能更多地涉及物理层以上的各层协议。

10.3.3 多点对多点系统

这种情况下，无线信道是多个节点共同使用的电磁介质。每个节点都可以使用这一介质不经访问基站而与其他任何节点直接交互而不需要访问基站。因为没有基站，所以需要使用分散算法来访问介质。

通常，使用这些组织无线信道的方法用来连接计算机（图10-9）。对于电话流量，在共享介质条件下获得带宽的不确定性显著降低了声音传输的质量。因此，电话网络总使用基站来分配带宽（即第3章的方法）。

第一个LAN于20世纪70年代在夏威夷建立。精确地对应于图10-9提供的设计。因为其通信速度很低，与当代使用的无线LAN不同，和低效的介质访问方法一起使用，导致只有18%的网络带宽利用率。

如今，这类网络在微波或红外线波段以高达52Mb/s的速率工作，全向天线用于多点对多点的通信。为了确保红外线的全向传播，使用扩散发送端（diffuse transmitter），它使用透镜系统分散光束。



图10-9 无线多点对多点信道

10.3.4 卫星系统

卫星通信用于建立高速、长距离的微波信道。这些信道需要视线通信。但却由于地球表面曲率而无法保证跨越很长的距离。卫星提供了解决这一问题的方法，并起到信号放射器的作用（图10-10）。

在苏联于1957年发射第一颗人造卫星之前，就已出现使用人造卫星创建通信信道的想法了。*Arthur C. Clarke*继续并进一步发展了*Jules Verne*和*H.G. Wells*的毕生工作，后两位发表了许多技术发明却未能实现。Clarke于1945年发表的题为《Extraterrestrial Relays》（宇宙中继）的论文中提出，在赤道上的地球同步卫星能保证地球上大部分区域的通信。

在冷战时期，由苏联发射的第一颗卫星不是可步卫星，且只能提供非常有限的通信能力。事实上，它只传输嘟嘟的无线电信号以告诉全世界它在外太空的出现。尽管如此，苏联的成功迫使美国迎头追赶。1962年，美国发射了第一颗通信卫星，名为Telstar-1，它可支持600路声音信道。

自世界上第一颗通信卫星发射成功40多年过去了，这些卫星的功能也自然变得更加复杂。如今，通信卫星可以扮演传输网络节点、电话交换机、计算机网络交换机或路由器的角色。现在，卫星设备可以与地面站和安装于其他卫星的设备交互，因此在太空中形成直接无线信道。在太空传输微波信号和在地球表面传输没有本质区别。尽管如此，卫星信道有其特性——其中一个节点形成了常在使用且与其他节点有明显距离的信道。



图10-10　卫星作为信号反射器

ITU为卫星通信分配了多个波段（表10-1）：

表10-1　ITU为卫星通信分配的频率波段

波段	下行频率（GHz）	上行频率（GHz）
L	1.5	1.6
S	1.9	2.2
C	3.7-4.2	5.925-6.425
Ku	11.7-12.2	14.0-14.5
Ka	17.7-21.7	27.5-30.5

C波段为首先使用的波段，在这一波段中，为每个地球-卫星（上行频率）和卫星-地球（下行频率）双工流分配500MHz，其足够用于组织数量庞大的信道。L波段和S波段也用于组建使用卫星的移动服务，通常，它们主要与在基于地面的系统中。K_u和K_a波段目前使用人数较少——因为昂贵的设备阻碍了大多数公司使用这一波段，尤以K_a波段为甚。

人造卫星按照*Johannes Kepler*提出的定律，绕地球运动，通常，卫星轨道为椭圆形，尽管如此，为了使卫星维持在地球表面上的固定高度，可选择圆形轨道。

如今，使用下列三种轨道，依据同地球表面距离的不同可分为（图10-11）：

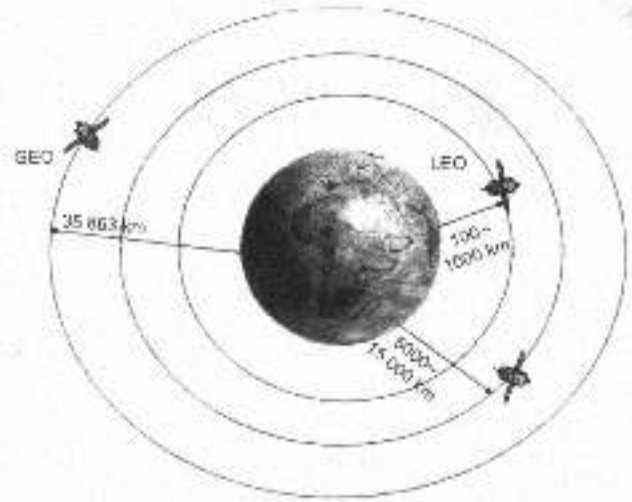


图　10-11

同步轨道，或GEO（35 863km）
中距离地球轨道，或MEO（5 000~15 000km）
近地轨道，或LEO（100~1 000km）

10.3.5 同步卫星

同步卫星（geostationary satellite）位于赤道上方的一个特定位置，绕地球旋转。这一位置方便通信，基于以下原因：第一，在任何时候地球的四分之一都在视线范围内，因此使用同步卫星，容易组建在整个国家甚至大洲内的广播。

第二，卫星对于建立在地球上的天线是固定的，这显著简化了组建通信的过程。事实上，这不再需要地面天线自动纠正方向（与MEO和LEO卫星相比）。这一情形随着20世纪90年代小型全向天线的产生而发生变化——现在，没有必要始终追踪LEO卫星的位置；它已足够保证使卫星在视线范围内。

第三，同步卫星轨道在地球大气之上，因此，损耗比MEO和LEO卫星小。LEO卫星受到大气阻力，重量持续减小，因此需用发动机周期性校正。

同步卫星通常以其拥有多天线而支持多信道。

通常，使用GEO卫星与使用直径超过10m的巨大天线相联系。这就增加了小机构和个人使用GEO卫星的复杂性。然而，安装在卫星上的定向天线的发展改变了这一情形。这些天线产生的信号可以用相对小的地面天线接收，称为甚小口径终端（VSAT）。VSAT直径大约为1m。装配了VSAT的地面站目前支持许多服务，诸如电话、数据传输和会议。

尽管如此，同步卫星难免有缺陷。最显著的就是其距离地球表面的距离太大。这导致明显的传输延迟，从230m到280m。当使用这种卫星传输对话或电话会议时，不方便的停顿发生干扰了正常的联系。

除这点外，在这一高度，信号损失率很高，这意味着有必要使用强劲的发送端和大直径天线。这与VSAT无关，但使用VSAT时，地面覆盖范围降低。

使用圆形轨道的地球同步卫星的根本缺点在于，靠近北极和南极地区时，通信质量不佳。比起赤道和中纬度地区，信号传输到这一地区需要跨越更长的距离。因此，有明显的衰减。解决这一问题的方法为发射一颗拥有明确椭圆轨道的卫星，其可以靠近地球南北极地区。

ITU也规范了同步卫星的轨道位置。如今，这些位置已有明显的缺点，因为各个GEO卫星在轨道上相互间不能小于2度。这意味着在每个轨道上，不能有超过180颗相似的卫星。不是每个国家都具有发射GEO卫星的能力，这一情形与特定频率波段的竞争类似。这种情况，又被竞争国家间的政治野心强化。

10.3.6 媒体和中低轨道卫星

目前，MEO卫星没有GEO和LEO卫星常用。MEO卫星确保覆盖直径范围为10 000到15 000km和50msec的传播延迟。由MEO卫星提供的最著名和最常用服务为**全球定位系统**（global positioning system, GPS）。GPS是一个用于确定用户在地球表面或接近地球的太空中当前坐标的全球系统。GPS由24颗卫星、一个用于追踪这些卫星的专用地面网络和不计其数的用户设备（接收端）构成。GPS持续（且相当精确）接收来自卫星的无线电信号，以确定用户当前位置的坐标。通常，误差从来不超过数十米。这对于实现移动物体（飞机、船只、汽车等）的导航工作已足够了。

LEO卫星比起同步卫星有不同的优点和缺点。这些卫星最大的优点就是接近地球。因此它们只需能量较低的发送端、小天线和较短的信号传输时间（20到25msec）。而且，它们易于发射。它们的主要缺点是覆盖范围小（直径只有8 000km）。这种卫星每隔1.5~2小时绕地球一周，且在基于地

面的站点可见间隔的时间很短——只需20分钟就可再次见到。这意味着只要发射了足够多的卫星就可以保证使用LEO卫星进行连续通信。除了这些以外，大气阻力把这类卫星的寿命减少至8~10年。

与主要用于广播和远距离固定通信的GEO卫星相比，LEO卫星被视为支持移动通信的主要方法。

20世纪90年代早期，摩托罗拉对LEO卫星设备的优点进行了评估。该公司与其他几个最大的伙伴合作，开始了铱星计划，其主要目标雄心勃勃——创建世界范围的卫星网络以确保任何位置的移动通信。到20世纪80年代末，移动电话的蜂窝系统并没有像现在这样密集。因此，保证了它的商业成功。

1997年，发射66颗卫星构成铱星系统。铱星系统的商业运作从1998年开始。铱星覆盖了地球的整个表面，在越过两极的六个轨道上运行。在每个轨道中，有11个装备有频率为1.6GHz和带宽为10MHz发送端的卫星。这一带宽用于组建240路每路41KHz的信道。归功于频率重用，铱星系统支持253 440条信道，因此建立了沿地球表面的蜂窝系统。提供给铱星用户的主要服务为电话通信（每分钟$7）和2.4Kb/s的数据传输。

铱星以其智能行为为特色。例如，它们可以使用专用的卫星间信道以25Mb/s交换数据。因此，从铱星电话上拨打的电话直接连接到目前视线内的另一卫星。这一卫星使用转发卫星系统将呼叫发送到目前离被呼叫用户最近的卫星。铱星系统是一个拥有确保有全球漫游功能的、完善的、专有协议栈的网络。

不幸的是，铱星的商业成功价值不大，存在两年后公司陷于破产。他们在移动电话用户发展上的投资是错误的。到他们的系统开始商业化运作时，基于地面系统的蜂窝电话已经在工业化国家覆盖。同时，确保2.4Kb/s的数据传输服务在20世纪末已经不能满足用户的需要。

今天，铱星系统再次投入运营。尽管如此，它拥有新的主人和新的商标名称——铱星卫星。此时，它已有了一个适度的计划，与在世界上还没有商业系统的地方创建用于移动通信的本地系统。卫星软件已升级且同时进行实用修改，允许数据传输率提高到10Kb/s。

全球星系统是另一个著名的LEO卫星网络。与铱星计划相比，48个LEO卫星的全球星系统实施传统的弯管功能。他们从移动用户接收电话呼叫，并把它传输给最近的地面基站。基站通过把来电传输给最近的卫星实施呼叫路由。如果被呼叫用户在其视线内。则不再使用卫星间信道。除了电话通信，全球星也提供4.8Kb/s的数据传输。

另一个LEO网络是商用低轨道小卫星通信系统（Orbcomm），它提供面向数据传输的服务。不幸的是，消息传输并不是实时模式。如果卫星不可见，那么商用低轨道小卫星通信系统的终端缓存数据分组，直到卫星再次出现在视野范围内。因此，数据传输为突发式。其延迟可达数分钟，而不是多数因特网用户所熟悉的延迟不超过一秒。

如今，移动网络主要由地面蜂窝网络支持，多数卫星系统改变了他们的目的。提供快速因特网访问变得前景光明。Teledesic即是这样的一个系统，它的发起者之一是微软的缔造者比尔·盖茨。这一系统于20世纪90年代早期开发，其中，卫星是由64Mb/s卫星间信道连接的全属性路由器。

基于GEO卫星的因特网访问系统——Spaceway、Astrolink和EuroSkyway——都在建设之中。这些系统面向于VSAT终端，并承诺为用户提供2~20Mb/s的访问信道。

10.4　扩频技术

扩频技术是专为无线传输开发的。它允许改善低能量信号编码的抗噪能力，这对于移动应用最重要。但要注意到，扩频技术并不是应用到微波波段无线信道的唯一编码技术。前面章节描述的任何种类的FSK和PSK都可使用。不使用ASK，不仅因为微波信道有广阔的带宽，而且还因为在宽频波段保证相同放大因子的放大器十分昂贵。

广阔的带宽也提供了使用多载波调制的可能性。在其中带宽被分为多个子信道，每个信道都用特定的载波频率。因此，比特流被分隔成多个子流，通常是第一个子载波频率（如f_c、$2f_c$、$3f_c$等）的几倍。使用标准的FSK或PSK方法调制。这一技术称为正交频分复用（orthogonal frequency division multiplexing, OFDM）。

传输前，所有的子载波使用快速傅里叶变换通过数学组织得到公共信号。这样信号的频谱大约等于使用单个载波进行编码的信号频谱。传输之后，使用反傅里叶变换检测载波子信道。然后，从每个信道接收比特流。将源高速比特流分为多个低速比特流，增加单独信号间的间隔。这意味着码间干扰作用由于电磁波的多路传播而降低。

10.4.1 跳频扩频

扩频的概念出现于第二次世界大战期间，当时无线电被广泛用于秘密谈判和军事目标控制，诸如鱼雷。为了防止监听或者用窄带噪声压塞无线电信号，通过在宽频范围内不停更换载波频率实施无线电传输，因此，信号能量在整个频率范围内分布。对某一特定频率的监听只能得到无意义的噪声。伪随机产生载波频率序列，只有发送端和接收端知晓。试图在特定频带内抑制信号不会对传播信号产生负面影响，因为只有一小份信息被抑制。

这一方法对应的概念，也称为跳频扩频技术（frequency-hopping spread spectrum, FHSS），如图10-12所示。

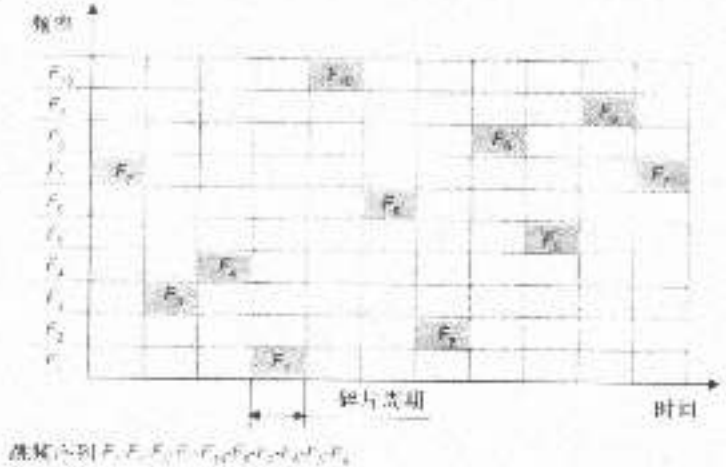


图10-12 跳频传播

以相同载波频率持续传输的时间间隔也称为碎片周期。在每个载波频率中，使用标准调制方法，如FSK或PSK。传输离散信息。为了将接收端和传输端同步，在每个碎片周期的开始，分配一个特殊时间间隔用于传输多个同步位的特殊时间间隔。因此，由于持续同步的系统开销，这一方法的有效速率较低。

载波频率根据伪随机数生成器产生的子信道和参数持续改变。伪随机序列依赖于称作种子的特殊参数，如果发送端和接收端都知道种子值，那么它们就可以根据相同序列改变频率。载波和率改变所依据的序列称作**频跳序列**（hopping sequence）。

子信道频率改变的频率称为碎片率（chipping rate）。如果碎片率低于信道信息率，这种操作模式称为慢速FHSS（slow FHSS）（图10-13a），而大多数情况下遇到的是快速FHSS（fast FHSS）（图10-13b）。

快速FHSS保证更好的抗噪能力，抑制特定子信道信号的窄带噪声不会导致位丢失，因为这一比特位值在不同子信道中重复多次，且只有其中的一份发生了失真。在这种模式下，不会发生码

间干扰，因为当信号到达时已经沿路由延迟，系统有时间切换到另一频率。

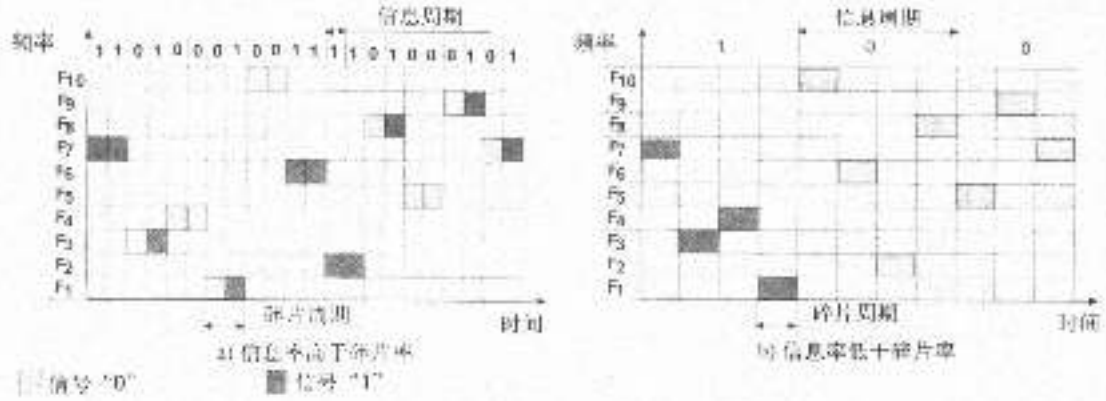


图10-13　信息率和碎片率之间的关系

慢速FHSS不能提供这一属性，即使它易于实现且系统开销较小。

FHSS技术应用于IEEE802.11和蓝牙等无线技术中。

FHSS方法所使用的频率波段与其他编码方法所采用的不同。不是为了节俭而使用狭窄的频率波段，而是试图占用整个可用的范围。起初，这种方法看起来不可能高效，因为在任意给定的时间内，只有一个信道在操作。但是，这是一个不精确的假设，因为扩频编码也在很广的范围内用于多路复用多个信道。特别是，FHSS允许通过为每个信道选择伪随机序列组建多信道操作，确保在任何时刻每个信道都以不同的频率工作。只有在信道数不超过可用子信道数时，可以实现这一点。

10.4.2　直接序列扩频

直接序列扩频（direct sequence spread spectrum, DSSS）方法也用于为无线链路分配整个频率波段。但其实现这一目的所采用的方法与FHSS所使用的方法不同。它使用分配给它的整个范围，但不使用持续变化的频率来实现。在这种情况下，每个信息位都用N个位替代以使信号传输时钟速率可以增加N倍，这也使得信号频谱增加了N倍。因此，适当选择信息率和N值足以使得信号频谱覆盖整个范围。

DSSS编码的目的与FHSS编码相同——也就是，改善抗噪性，窄带噪声只能使信号频谱的特定频率失真。因此，接收端可以正确识别出传输信息的可能性极高。

用于替换源信息一个位的编码称为扩展序列，且这种序列的每位称为碎片（chip）。因此，结果编码的传输率称为碎片率。二进制0使用扩展序列的相反值编码。接收端必须知道发送端所使用的扩频序列才能正确识别出传输信息。

扩展序列中位的数目是扩展因子，因为其定义了源码的扩充系数。与FHSS相同，可以使用任何调制方法（例如，BFSK）编码结果码。

扩展因子值越大，结果信号的频谱越宽，其噪声抑制级别也越高。而信号所占用的频谱波段也增大。通常，扩展因子取10到100之间的值。

扩展序列的一个实例为Barker序列，其由11位构成：10110111000。如果发送端使用这一序列，那么传输二位110将会生成以下位进行传输：

10110111000 10110111000 01001000111

Barker序列允许接收端快速完成与发送端的符号同步（也即，可靠地检测序列的开始点）。接收端将接收到的位与序列模块按顺序进行比较。如果Barker序列与相同模式相比右移或左移了一位，那么不到一半的位可以匹配。这也意味着即使有几位失真了，接收端仍然会以很高的概率正确决

定序列的开始位置。因此，其可以正确解释接收到的信息。

DSSS没有FHSS抗噪能力强劲，因为高强度窄带噪声可以影响部分频谱，因此也影响了0和1的识别。

10.4.3 码分多路访问

与FHSS相同，DSSS编码允许在相同波段内多路复用多个信道。这种多路复用技术称为**码分多路访问**（code division multiplexing access, CDMA），它被广泛应用于蜂窝网络中。CDMA多路复用技术可以使用FHSS和DSSS编码。实际上，它频繁用于使用DSSS码的无线网络中。

使用CDMA网络的节点在任何需要的时候发送数据给共享介质，这意味着在网络节点间没有同步机制。CDMA的基本思想是，每个网络节点使用自己的扩展序列。选择序列值来保证接收端识别发送端的扩展序列，从信号中拾获发送节点创建的数据，而信号是多个节点同时传输的结果。

为了确保可以实现这样的多路复用，以特定方式选取扩展序列的值。让我们用实例来解释CDMA。

假设网络中有4个节点：*A*、*B*、*C*和*D*。每个节点使用自己的扩展序列：

A：0 0 0 0

B：0 1 0 1

C：0 0 1 1

D：0 1 1 0

还假设用扩展序列（传输源码）传输0、1时，使用加性信号和反转信号。反转属性意味着对于二进制1，使用+A振幅的正弦波编码；二进制0使用-*A*振幅的正弦波编码。对于加性属性，其遵循当同时传输二进制1和0时，如果这些正弦波的相位相符，那么信号级别为0。为了简化传输序列的书写，指定正弦波的正振幅为+1，负振幅为−1。简单起见，假设所有的CDMA网络节点都是同步的。

因此，当传输源码1时，四个节点将如下序列传入介质：

A：−1 −1 −1 −1

B：−1 +1 −1 +1

C：−1 −1 +1 +1

D：−1 +1 +1 −1

当传输源码0时，扩展序列的信号相反。

现在，假设每个节点与其他节点独立地传输源信息中的一位：*A*节点传输位设置为1，*B*节点设置为0，*C*节点设置为0，*D*节点设置为1。

以下是传输到网络介质（S）中的信号序列：

A：−1 −1 −1 −1

B：+1 −1 +1 −1

C：+1 +1 −1 −1

D：−1 +1 +1 −1

根据加性属性：

S：0 0 0 −4

假设*E*节点需要从*A*节点接收信息。为了实现这一目的，需要使用以节点*A*的扩展序列作为参数来指定的CDMA解调器。

CDMA解调器以如下方式工作：它在每个时钟的操作中逐个把接收到的四个信号（S_i）相加。当*A*节点的扩展码为+1时，此刻接收到的S_i连带符号一起考虑，而当*A*节点的扩展码为−1时，接收

到的S_i则以反转符号形式加入到和中。换句话说，解调器计算信号向量和所需节点扩展序列向量的标量积：

$$S \times A = (0\ 0\ 0\ -4) \times (-1\ -1\ -1\ -1) = 4$$

为了找出A节点发送的位，结果需要标准化，例如，通过除以网络中的节点数：4/4 = 1。

如果E节点要接收来自B节点的信息，需要使用该节点用于解调的扩展码，或$B(-1\ +1\ -1\ +1)$：

$$S \times B = (0\ 0\ 0\ -4) \times (-1\ +1\ -1\ +1) = -4$$

标准化后，我们得到信号−1，其对应于B节点源信息中的二进制0。

使用CDMA的扩展序列，其特色为序列之间相互正交。这意味着如果把它们视为向量，两两相乘得到0。三维空间向量（1 0 0），（0 1 0）和（0 0 1）是另一个相互正交向量的实例。尽管如此，使用CDMA中的向量，不仅要保证相互之间正交，也要保证和向量集合中的其他反转成员正交。这是由于反转数用于编码源信息中的0。

这里，我们只用尽可能简单的方法解释CDMA的基本思想。实际上，CDMA是不对传统的+1和−1进行操作的复杂技术；更精确地说，它对调制信号进行操作，例如BPSK信号。此外，网络节点也是不同步的。最后，从位于不同位置不同距离的节点到达的信号拥有不同的能量。接收端和发送端间的同步问题，可以通过发送事先定义的编码，称为领航信号（*pilot signal*）的长序列来解决。为了使所有的发送端对于基站而言能量大致相同，CDMA使用特殊的过程用于控制能量。

小结

- 无线通信分为移动无线和固定无线两种。为了组织移动通信，没有可替换的无线介质。固定无线通信确保对坐落于限制地域中的网络节点的访问。
- 无线通信信道的每个节点都装配有天线，其同时为电磁波的发送端和接收端。
- 电磁波可以在所有方向（全向）或某个扇区内（有向）传播。传播类型取决于天线的类型。
- 无线数据传输系统根据电磁波频谱的范围分为以下四种：
  - 无线电（广播）系统
  - 微波系统
  - 红外线波系统
  - 可见光系统
- 由于电磁波的反射、衍射和散射，会发生相同信号的多路传播。这将导致符号间的干扰和多路衰减。
- 以900MHz、2.4GHz和5GHz波段进行的数据传输称为工业、科学、医学波段，不需要许可，但传输能量不能超过1W。
- 无线点对点信道用于创建无线电中继线路，实现建筑物间和计算机间的通信。
- 无线点对多点信道创建在基站的基础上。这些信道用在移动蜂窝网和固定访问系统中。
- 多点对多点拓扑是无线LAN的特色。
- 卫星通信系统使用三组圆轨道，依照其与地球表面距离的不同分为：
  - 同步轨道（35 863km）
  - 中地轨道（5 000~15 000km）
  - 近地轨道（100~1 000km）
- 为了编码离散信息，无线系统使用下列的调制方法：
  - FSK和PSK
  - 使用多个载波频率的OFDM调制

- 扩频方法——跳频扩频（FHSS）和直接序列扩频（DSSS）
- 扩频方法使用频率范围表示信息。这降低了窄带噪声对信号的影响。
- 基于FHSS和DSSS，可以在相同频率波段内复用多个信道。这一多路复用技术称做码分多路访问。

复习题

1. 列出无线通信信道的主要应用领域。
2. 与使用导向介质的传输方法相比，无线信息传输有什么优点和缺点？
3. 怎样可以组织无线电波和微波的全向传播？
4. 什么因素允许频率范围从2MHz到30MHz的无线电波传输数百公里？
5. 什么频谱用于卫星通信？
6. 哪种大气条件阻止了微波传输？
7. 什么设备用于红外线波的全向传输？
8. 什么障碍物导致电磁波的衍射？什么导致散射？
9. 什么时候需要使用通信卫星的椭圆轨道？
10. 同步卫星的缺点是什么？
11. 在你看来，铱星计划商业失败的原因是什么？
12. 若要FHSS技术的传播速度加快需要遵循什么条件？
13. Barker序列的什么属性使其用于DSSS技术中？
14. 用于CDMA扩展序列的主要属性是什么？

练习题

1. 是否可以使用1 0 0…0，0 1 0 0…0，0 0 1 0…0，0 0 0 1 0…0等作为支持基于DSSS技术的CDMA网络节点的扩展序列？
2. 给出一个不同于Barker序列的11位扩展序列，并提供检测传输源信息下一比特开始实例的可靠性。

第11章 传输网络

11.1 引言

传输网络（transmission networks）用于创建交换基础结构，使用它可在两个端用户设备间组建快速、灵活的永久的“点对点”信道。传输网络采用电路交换技术。覆盖网络，例如计算机或电话网络，在由传输网络形成的电路上运作。由传输网络提供给用户的信道以高带宽而闻名——通常从2Mb/s到10Gb/s。

三代传输网络：

- 准同步数字系列（PDH）
- 同步数字系列（SDH，在美国，SONET标准对应于SDH技术）
- 密集波分复用（DWDM）

前两种技术——PDH和SDH——使用频分复用共享宽带链路且以数字形式传输数据。每个技术都支持传输率层次，所以用户可以在所建的覆盖网络上为信道选择所需的速率。

SDH技术保证了比PDH 更高的传输率；因此，当建立一个大型的传输网络时，它的主干网通常基于SDH技术，而访问网络采用PDH技术。

DWDM是创建快速通信信道的最新成果。它并非数字形式，因为它提供给用户单独的电波用于信息传输。用户可以随心所欲地使用，实现调制或编码。如今，DWDM技术已迫使SDH方法退出了长距离主干网而进入网络外围，将SDH转变为访问网络技术。

交换和多路复用的三种技术允许创建灵活和可升级的、能够为大多数计算机和电话网络服务的传输网络。

11.2 PDH网

准同步数字系列（plesiochronous digital hierarchy，PDH）技术在20世纪60年代由AT&T在解决大型电话网络交换机互联问题时开发出来。之前使用FDM信道解决这一问题，那时其已不堪重负来通过单一电缆组建高速多信道通信。在FDM中，双绞线用于在12路本地环路上同步传输数据。为了改善通信速率，有必要安装数对电缆或者使用更加昂贵的同轴电缆替代。

11.2.1 速率层次

PDH技术革命的起点是T-1多路复用器的开发，它可以用数字形式（连续）为24个用户多路复用、交换和传输声音。因为用户继续使用标准电话机，这意味着以模拟形式传输声音，T-1多路复用器自动以8 000Hz速率采样声音，且使用脉冲编码调制来编码声音。因此，每个本地环路形成64Kb/s数字数据流，且整个T-1线路保证1.544Mb/s[⊖]的带宽。

对于连接大型、自动的数据交换机，T-1链路并不是足够强劲和灵活的多路复用工具。因此，考虑实现具有**速率层次**（rate hierarchy）的通信链路思想。四个（*four*）T-1链路合并形成下一级数字层次——T-2，它以6.312Mb/s的速率传输数据。通过合并七个T-2链路得到44.736Mb/s速率的T-3链路。合并六个T-3链路得到T-4链路；因此，它的传输率为274.176Mb/s。这一技术称为**T载波系统**（T-carrier system）。

⊖ 原书为1,544Mb/s，应为1.544Mb/s，本小节中速率值均有此错误。——译者注

在20世纪70年代中期，电话公司开始提供建立在T载波系统上的专用链路用于商业租赁，且不再是公司内部技术。T-1——T-4链路不仅可以传输声音，而且还可以传输任何以数字形式表示的数据，包括计算机数据、电视和传真。

T载波系统由美国国家标准协会以及后来被CCITT，即现在的ITU-T，标准化。标准化后，称为PDH。由于CCITT的更正，美国和国际版本的PDH标准变得不兼容。在国际标准中，T链路的对应为E-1、E-2、E-3和E-4链路，且速率分别为2.048Mb/s、8.488Mb/s、34.368Mb/s和139.264Mb/s。美国版本的标准也在加拿大和日本（略有不同）采用；在欧洲，使用国际CCITT标准。

虽然美国标准和国际标准不同，但数字系列技术在指定速率层次时使用相同的符号——**数字信号级n**（digital signal level n, DS-n）。表11-1列出了两个技术标准下所有级别的数据传输率值。

表11-1 数字数据率的层次

	美 国			CCITT（欧洲）		
速率名称	声音信道的数量	前一层次级信道的数量	速率，默认单位Mb/s	声音信道的数量	前一层次级信道的数量	速率，默认单位Mb/s
DS-0	1	1	64Kb/s	1	1	64Kb/s
DS-1	24	24	1.544[⊖]	30	30	2.048
DS-2	96	4	6.312	120	4	8.488
DS-3	672	7	44.736	480	4	34.368
DS-4	4 032	6	274.176	1 920	4	139.264

实际中，*T-1/E-1*和*T-3/E-3*最常用。

11.2.2 多路复用方法

T-1多路复用器确保在一个简单格式的帧中以1.544Mb/s传输来自24个用户的数据。在该帧中，用户数据按顺序传输，每个用户每次传输一个字节，24个字节后添加一个同步位（*synchronization bit*）。最初，T-1设备（通称整个以1.544Mb/s工作的数据传输技术）只在内部时钟下工作，且每个帧可以使用同步位进行异步传输。T-1和更快的T-2及T-3设备在随后的很长时间内发生了明显的改变。如今，传输网络的多路复用器和交换机已经使用由一个固定网络位置发布的中央时钟来运行。但帧的形成规则保持不变；因此，同步位依然存在于帧中。用户速率的总速率为24用户 × 64Kb/s = 1.536Mb/s，并同步位添加的8Kb/s，为1.544Mb/s。

现在，考虑T-1帧格式的另一个特点。在T-1设备中，每个帧每个字节（*each byte*）的第8位根据传输数据类型和产生设备有特殊的含义。当传输声音时，这一位用作服务信息，诸如呼叫用户数和网络用户间建立连接所需的其他信息。在电话学中确保这种连接的协议称为**信令协议**（signaling protocol）。因此，用户声音传输的实际速率为56Kb/s而不是64Kb/s。使用第8位用于服务目的称为**位劫持**（bit robbing）。

当只传输计算机数据时，T-1线路提供23条信道用于用户数据，第24条信道专用于内部使用，主要重储损坏的帧。计算机数据以64Kb/s传输，因为第8位没有被“窃取”。

在同时传输声音和计算机数据时，使用所有的24路信道，且声音和计算机数据都以56Kb/s速率传输。

当将四路T-1信道多路复用为单个T-2信道时，同前面一样，在DS-1帧之间添加同步位，而DS-2帧（包含四个DS-1帧）由12个服务位分割，它不仅用于分割帧还用于同步帧。因此DS-3帧由服务位分隔的七个DS-2构成。

⊖ 原书中为1,544，应为1.544，表格中速率值均有此错误。——译者注

前面已经提到，在CCITT国际标准G.700-G.706中的PDH技术版本与美国T载波系统技术不同。更确切地说，它不使用位劫持方法。当到达下一级层次时，速率乘数有恒定值4。E-1线路不使用第8位，而是从32字节中分配2字节用于服务目的，即，第0字节（用于发送端和接收端间的同步）和第16字节（用于传输信令信息）。对于声音或数字传输，可用30条信道，每条信道的速率为64Kb/s。

用户可以租赁T-1/E-1链路的多条56/64Kb/s信道。这种复合信道称为分割*T*-1/*E*-1（*fractional T*-1/*E*-1）。在这种情况下，多路复用操作为用户分配多个时间槽。

PDH技术物理层支持多种类型的电缆：双绞线，同轴电缆和光纤电缆。组建用户访问T-1/E-1链路的主要变化形式为用RJ-48插座连接的两条双绞线构成的一条电缆。需要两对双绞线用于组建以速率1.544/2.048Mb/s进行数据传输的双工方式。T-1信道使用B8ZS双极电平编码表示数据，而E-1使用HDB3双极电平编码实现相同目的。对于T-1线路的信号放大，再生器和线路控制设备安装的间隔为1 800m（1英里）。

同轴电缆，由于它的高带宽，支持T-2/E-2信道或者四条T-1/E-1信道。对于T-3/E-3信道，则使用同轴电缆光纤电缆或者微波信道。

在G.703标准中描述了这一技术国际变化形式的物理层，其名字指定了连接到E-1信道的桥或路由接口的类型。标准的美国版本的名字为T-1。

11.2.3　PDH技术的局限性

美国和国际版本的PDH技术标准都有许多缺点，其中最重要的是复杂性和针对用户数据多路复用/解多路复用操作的低效性。命名这一技术所使用的术语本身——准同步（也即接近同步）——隐含地证明了在将慢速信道多路复用为快速信道时缺乏完全的同步。最初，用异步方法传输帧时，需要在帧之间插入同步位。

因此，为了从多路复用信道上取回用户数据，需要从聚合信道上**完全**（fully）分离（解多路复用）帧。例如，为了从T-3信道中取回特定64Kb/s用户信道的数据，需要把这些帧分离成T-2级的帧，继续分离为T-1级帧，最终分离为T-1帧。

如果PDH网络只用作在两个大型节点间建立主干网，那么多路复用/解多路复用操作只在端节点上实施；因此，不会产生问题。但是需要将PDH网络的传输节点上的多个用户信道进行分解时，就没有简单的解决方法。解决这一问题的可能方法是在每个网络节点上安装两条T3/E3（或更高层）多路复用器（图11-1）。第一个多路复用器携带数据流的完全解多路复用结果，且把慢速信道定向到用户；第二个设备再次把其余信道多路复用为高速输出流。当实施这一方法时，操作的设备数翻倍增长。

另一种变化形式为**网络回程**（back hauling）。转发节点需要分解和重定向用户数据流，因此只安装高速解多路复用器，它只是将数据在网络中传输而不对数据进行解多路复用操作。只有端节点的解多路复用器实施解多器复用操作，此后特定用户的数据通过独立的物理链路返回到传输节点。自然，这些繁琐的交换机间的操作使整个网络操作复杂并且需要调整。这增加了所需的人工配置操作的数，并导致错误发生。

此外，PDH技术不提供内置网络管理和容错方法。

最终，PDH只能提供比当前需求慢的信息传输率。光纤电缆允许在单一信道上以数吉比特每秒的速率传输信息，这可使数万用户信道多路复用一根电缆。但是，PDH技术不具有这一能力，因为其层次速率限制在139Mb/s。

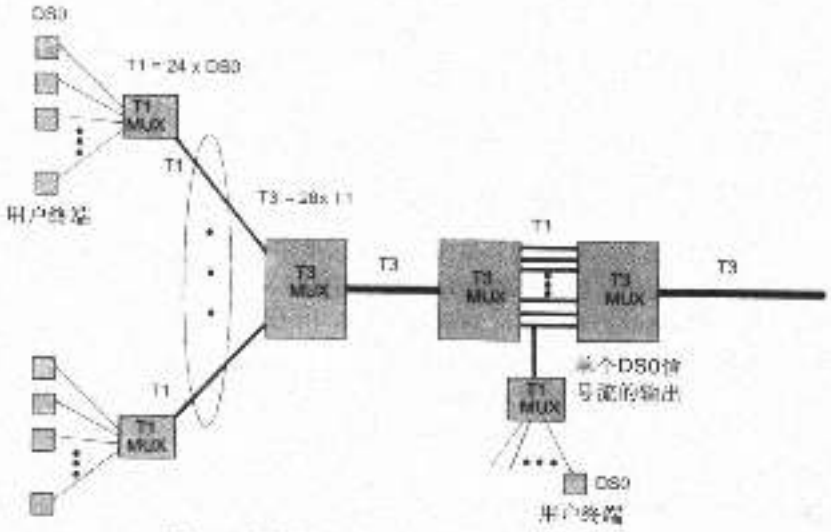


图11-1 使用完全解多路复用的低速信道的分离

11.3 SONET/SDH网

同步光网（synchronous optical NET, SONET）的设计人员考虑到并消除了前面提到的PDH技术的缺点。这一标准首先出现在1984年。此后，这一技术由ANSI委员会标准化。在欧洲电信标准委员会（ETSI）和ITU-T的协调及ANSI、美国、欧洲、日本的顶级电信公司合作下，出台了这项技术的国际标准。开发国际标准的目的是为了创建能够使用基于光纤的高速主干网传输现存PDH级别（美国T1-T3和欧洲E1-E4）数字信道流量的技术，并保证可以继续使用PDH层次达每秒几个吉比特的速度层次。

由于长期合作，ITU-T和ETSI开发出称为**同步数字系列**（synchronous digital hierarchy, SDH）的国际标准。SONET标准也被证实可以确保设备的兼容性。因此，SDH和SONET网络是兼容的（但不相同）且可以多路复用实际中任何PDH标准的输入流——无论是欧洲的还是美国的。

11.3.1 速率层次与多路复用方法

由SONET/SDH技术支持的速率层次如表11-2所示。

表11-2 SONET/SDH速率层次

SDH	SONET	速 率
	STS-1, OC-1	51.84Mb/s
STM-1	STS-3, OC-3	155.520Mb/s
STM-3	OC-9	466.560Mb/s
STM-4	OC-12	622.080Mb/s
STM-6	OC-18	933.120Mb/s
STM-8	OC-24	1.244Gb/s
STM-12	OC-36	1.866Gb/s
STM-16	OC-48	2.488Gb/s
STM-64	OC-192	9.953Gb/s
STM-256	OC-768	39.81Gb/s

在SDH标准中，所有的速率级（以及这些级别的帧格式）都有适用名称：同步传输模式级N

(STM-N)。在SONET技术中，有两个速率级的名称：**同步传输信号级N**（STS-N），用于以电信号形式进行的数据传输，和**光载波级N**（OC-N），在当数据由光纤电缆传输数据时使用。后面，为简单起见，我们集中讨论SDH技术。

STM-N帧具有的复杂结构将SDH和PDH以不同速率整合到公共主干流中，且可以不用对骨干流过行解多路复用就可实现添加/丢弃操作。

多路复用和添加/丢弃操作使用虚容器（virtual container），PDH类可在SDH网络上传输。除了PDH数据块，辅助控制信息也添加到虚容器中，包括容器路径的系统开销报头。这一报头包含静态信息，沿着容器传输过程中从源到目的地的路由（出错报文），和诸如端点间连接建立指示器的控制数据等。因此，虚容器的容量通常比它承载的对应于PDH的有效负载要大。例如，VC-12，除了32字节的E-1流数据外，还包含3字节的控制信息。

在SDH技术中，有多种类型的虚容器（图11-2）用于传输PDH数据块，其主要类型为：VC-11（1.5Mb/s），VC-12（2Mb/s），VC-2（6Mb/s），VC-3（34/45Mb/s）和VC-4（140Mb/s）。

虚容器是SDH多路复用器的交换单元。在每个多路复用器中，都有**连接表**（connection table），也称为**交叉连接表**（cross-connection table），它规定了，例如，P1端口的VC-12连接到P5端口的VC-12，以及P8端口的VC-3连接到P9端口的VC-3。网络管理员使用管理系统或者使用控制终端在每个多路复用器上产生连接表。这一操作必须以这样的方式进行，以确保用户设备连接到的两个网络端点间的连接路径。

为了将同一网络中的STM-N帧的同步传输和由这些帧承载的PDH用户信息的异步属性相结合，SDH使用指针（point）技术。指针的概念是SDH技术的关键，它使用冗余位替代了在PDA技术中使用的异步源速率均衡。

指针决定了在更高层结构中虚容器的当前位置——**从属单元**（tributary unit）或**管理单元**（administrative unit）。这些单元与虚容器的主要差别是额外指针域的存在。因为指针的使用，虚容器可以在从属单元或管理单元内的一定区域内“漂移”，这些单元相对而言在帧中有固定的位置。由于指针系统，多路复用器能够找到STM-N帧同步字节流中用户数据的位置，并在“传输中”将它们取回。注意，在PDH中实现的多路复用机制没有提供这一能力。

从属单元结合成组，这些组再组成管理单元。*N*个管理单元的组组成STM-N帧的有效载荷。此外，帧含有一个对于所有管理单元通用的报头，它包含控制信息。在用户数据传输的每一步中，将多个控制字节添加到源数据中。这些辅助字节有助于识别块结构或块组，然后使用指针决定用户数据的开始位置。

图11-2显示了SDH帧的结构单元。包含指针的单元用阴影标出。允许数据位置变化的容器和单元间的关系用虚线表示。

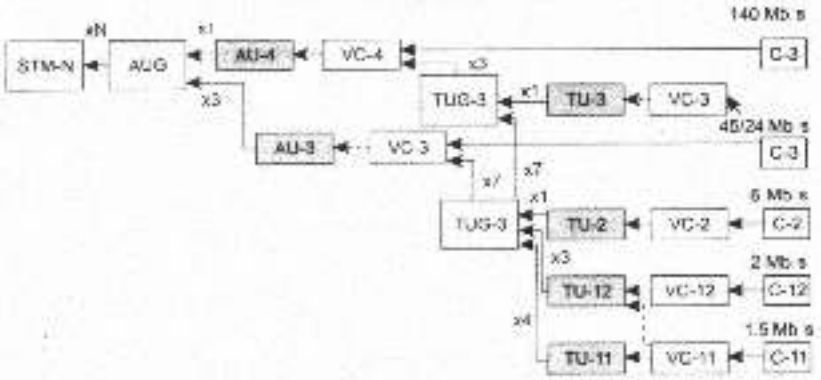


图11-2　SDH中的数据多路复用方法

SDH多路复用方法提供了聚合PDH用户流的不同能力，例如，对于STM-1帧，可以实现以下变化形式：

- 1 E-4流
- 63 E-1流
- 1 E-3流和42 E-1流

当然，你可以使用其他变化形式。

11.3.2 设备类型

SDH网络的主要元素为**多路复用器**（multiplexer）（图11-3）。通常它装配有一定数量的PDH和SDH端口，例如，2Mb/s和34/45Mb/s PDH端口，用于155Mb/s的STM-1端口及用于622Mb/s的STM-4端口。SDH多路复用端口分为聚合端口和从属端口。*从属端口*（*tributary port*）常称*添加/去弃端口*（*add/drop port*），而*聚合端口*（*aggregate port*）也称为*线路端口*（*line port*）。这一术语也反映了SDH网络的典型拓扑结构，这里有明确的链或环形式的主干（线路），沿着这些线路来自网络用户的数据流通过添加/去弃端口传输。

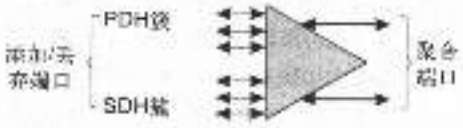


图11-3 SDH多路复用器

SDH多路复用器通常分为下列两种类型，两者间的区别取决于它们在SDH网络中多路复用器的位置。

- **终端多路复用器**（terminal multiplexer）。终端多路复用器通过多路复用多条从属信道来结束聚合信道（图11-4A）。因此，终端多路复用器装配有一个聚合端口和多个支路端口。
- **添加/丢弃多路复用器**（add/drop multiplexer）。添加/丢弃多路复用器在主干网（环、链、或混合拓扑）中占据中间的位置，它有两个聚合端口用于传输聚合数据流（图11-4B）。使用少量的从属端口，该多路复用器可以向聚合数据流增加支路信道的数据或减少聚合数据流的支路信道数据。

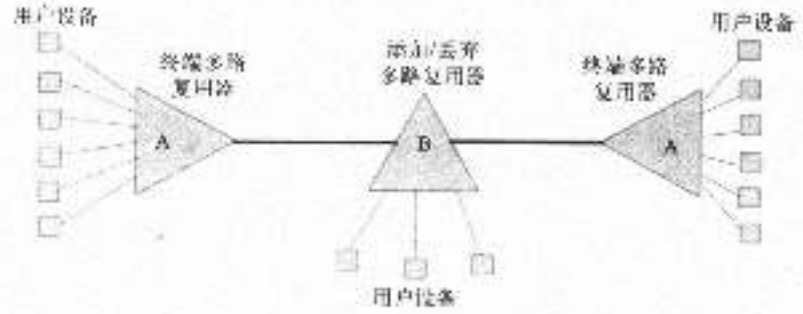


图11-4 SDH多路复用器的类型

有时使用**数字交叉连接**（digital cross-connect）——在任意虚容器上实施交换操作的多路复用器。在这种的多路复用器中，聚合端口和从属端口没有差别，因为这种多路复用器在网状拓扑中使用，这种情况下不可能选出聚合流。

除了多路复用器以外，SDH网络还可以包含必要的**再生器**（regenerator）来消除多路复用器间距离的限制，这些限制取决于光发送端的能量、接收端的敏感度和光纤电缆的衰减（这一问题已经在第8章中的能量预算的练习讨论过了）。再生器将光信号转变为电信号，然后再转变回光信号，在这操作过程中，重新存储信号波形和时间特性。目前，很少使用SDH再生器，因为它的成本并不比多路复用器有明显降低，而功能却不能和多路复用器相比。

11.3.3 协议栈

SDH协议栈（SDH protocol stack）由四层协议构成。这些层与OSI模型的层无关。对于OSI模型，整个SDH网络由物理层设备构成。

- 在SDH协议栈中，光层（photonic layer）使用光调制处理信息位的编码。为了编码光信号，使用NRZ电平码。并在传输前通过对数据扰频实现编码的自同步属性。
- 段层（section layer）支持网络的物理层整合。在SDH术语中，段指连接一对SONET/SDH设备（例如，可能连接多路复用器与再生器或者再生器与另一个再生器）的光纤电缆的连续部分。段也常称为**再生器段**（regenerator section）。我们假设实现这一层实现的功能不需要段的终结设备进行多路复用。再生器段协议用于处理帧头的特定部分，称为**再生段开销**（regenerator section overhead）。在控制信息的基础上，这一协议可以测试段机支持管理控制功能。
- **线路层**（line layer）负责在两个网络的多路复用器间传输数据。这层协议为多路复用、解多路复用添加删除用户数据来处理STS-N层帧。如果其中的元素失效——可能是光纤、端口或者相邻的多路复用器线路协议也用来负责线路识别。线路也常称为**多路复用段**（multiplex section）。多路复用段的控制信息也置于称为**多路复用器段头**（multiplex section overhead, MSOH）的帧头部。
- **路径层**（path layer）负责在网络的两用户间进行数据传递。路径是用户间的组合虚连接。路径层协议必须接收以用户格式（例如，以T-1格式）提供的数据，并将它们转化为同步STM-N帧。

图11-5通过SDH设备的类型说明了SDH协议的分布。

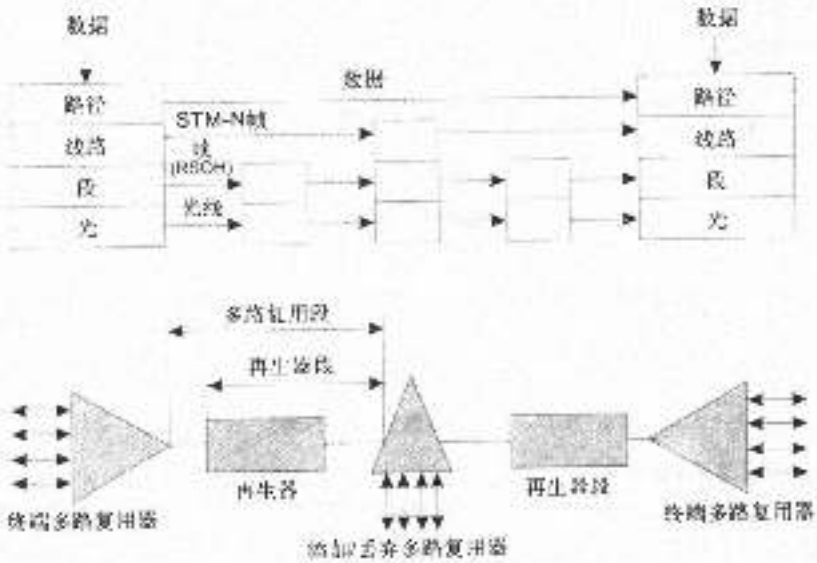


图11-5　SDH技术的协议栈

11.3.4 STM-N帧

STM-1帧的主要元素如图11-6所示。表11-3列出了再生器和多路复用段的头部结构。通常，STM-1帧由270列、9行构成的矩阵表示。每一行的前9个字节用于头部的服务数据，接下来261字节中的260字节为有效负载（例如，管理单元、管理单元组、支路单元、支路单元组和虚容器等数据结构）。每行的一个字节作为路径头，它控制点到点的连接。

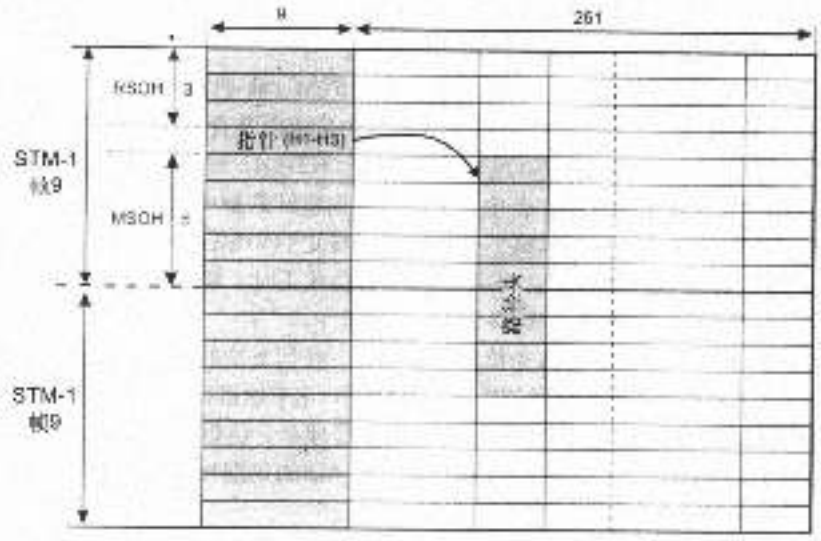


图11-6 STM-1帧的结构

表11-3 再生器和多路复用段头的结构

再生器段头部	多路复用段头部
同步字节 再生器段的错误控制字节 一个字节的辅助无线电信道 3字节的数据通信信道（DCC），以192Kb/s的速率运作 国家通信数据保留字节 指针域的H1，H2和H3指定了与指针域相关的VC-4或者3个VC-3的起始位置	用于多路复用段的错误控制字节 6个字节的DCC，以576Kb/s速率运行 2字节的自动流量保护协议（第K1和K2字节），确保网络存活 1个字节的同步系统状态报文 MSOH头的所有剩余字节又留用于国家通信或者不使用

考虑到H1-H2-H3指针操作机制，以STM-1帧为例，承载了VC-4。指针占用帧的第四行的第九个字节，给每个H1，H2和H3域都分配3字节。允许的指针值在0到782之间。指针用3字节单元标记VC-4的开始位置。例如，如果指针设为27，那么VC-4的第一个字节将位于指针域的最后一个字节的$27\times3=81$字节处，这意味着它是STM-1帧第四行的第90个字节（从1开始计数）。指针的固定值考虑到了数据源（其角色可由PDH多路复用器、拥有PDH/SDH接口的端用户设备或者其他的SDH多路复用器来扮演）和当前多路复用器间的相位移动问题。因此，用两个连续的STM-1帧传输虚容器，如图11-6所示。

指针不仅可以处理固定相位移动也可以处理多路复用器和接收用户数据的设备间的时钟频率不匹配问题。为了补偿这一作用，指针值周期性增加或者减少1。

如果VC-4的数据到达速率低于STM-1的发送速率，多路复用器将会周期性（这一周期取决于同步频率不匹配的值）经历用户数据短缺。这些用户数据用来填充虚容器的适当域。因此，多路复用器插入三个“哑”（无意义的）字节到虚容器数据中，此后连续用等额期间接收到的用户数据填充VC-4。指针增加1代表下一个VC-4的开始位置移动了3个字节。对指针的这一操作称为正校正（positive alignment）。因此，用户数据传输的平均速率等于在PDH方式不插入无会位的数据到达速率。

如果VC-4数据到达的速率高于STM-1帧发送的速率，那么多路复用器需要周期性地在帧中插入“额外”字节（额外指VC-4域中没有这些字节的位置）。为了放置这些字节，使用指针的三个最

不重要的字节——即，H3域（指针值本身在H1和H2域的字节中填充）。在这种情况下，指针值减少1；因此，这一操作称为负校正（negative alignment）。

可以简单地解释VC-4校正与用3个字节单元的原因。STM-1帧可以承载一个VC-4或者三个VC-3。一般来说，每个VC-3拥有与帧开始位置相关的独立相位值以及自己不匹配的频率值。与VC-4指针不同，VC-3指针由3个字节构成而不是9个字节：H1、H2和H3（这些域均为1字节长），像VC-4指针一样，这三个指针放置于相同的字节中；然而，根据指针的顺序：H1-1、H1-2、H1-3、H2-1、H2-2、H2-3、H3-1、H3-2和H3-3（第二个索引指示特定的VC-3），使用交叉方法，用字节而不是以3字节单元形式表示VC-3指针值。当完成VC-3的负校正时，额外字节被放置在适当的字节中——H3-1、H3-2或者H3-3——取决于对VC-3所实施的操作。

因此，我们必须说明VC-4偏移量的选择。选择它统一任何类型的容器上的操作，将它直接放置到STM-1域的管理单元组中。低层容器的校正总是1字节1字节地进行。

当根据以上描述的方法（图11-6）将从属单元块和管理单元块整合成组时，他们依次逐字节交叉，以便到达STM-N帧中的用户数据的周期与到达从属端的周期相等。这就排除了对临时缓存的需要；因此，可以说*SDH*多路复用以实时方式传输数据。

11.3.5　典型的拓扑

SDH网络中使用不同的链路拓扑。最常用的为环和多路复用器线性链。网状拓扑近似于全连结，它的应用领域也在不断增长。

SDH环（SDH ring）建立在不少含有两个聚合端口的添加/丢弃多路复用器基础上（图11-7a）。使用从属端口将用户数据流添加到环中，并从环中丢弃，构成点对点通信连接（图中显示了两个这样的连接）。环是典型的拥有容错能力的规则拓扑——在一根电缆损坏或者一个多路复用器失效的条件下，仍然能够保持连接，如果它直接沿着环的相反方向传输。环通常建立在两条光缆之上。有时，用四条光纤电缆改善可靠性和带宽。

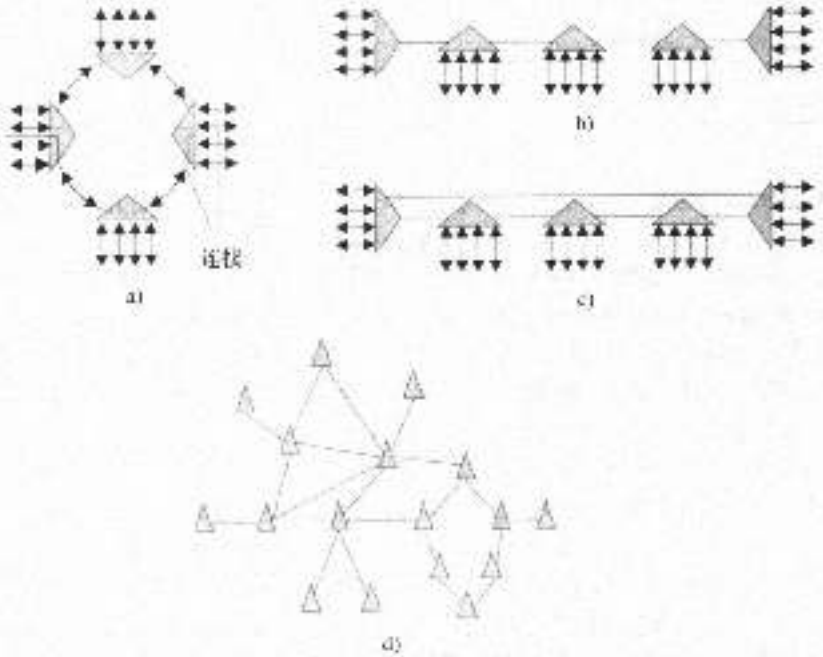


图11-7　典型的拓扑

链（图11-7b）是多路复用器线性序列（linear sequence of multiplex），其中两端的多路复用器起到终端的作用，而其他的多路复用器则为添加/丢弃多路复用器。通常，拥有链拓扑的网络使用在特殊地理位置的地方，例如，当网络不得不沿铁路和土下线管道分布时。尽管这样，在这些情况下仍可以使用平面环（图11-7c）。通过在主干电缆中使用两个额外的光纤和在每个终端多路复用器上额外聚合端口，确保了更高级别的容错。

这些基本拓扑可以使用多个分支用辐射环拓扑创建段和"环到环"连接等构成一个复杂而可扩展的SDH网络。最通用的情况为网状拓扑（图11-7d），在这里，多路复用器通过许多链路彼此相连，确保达到高层的性能和可靠性。

11.3.6 保证网络抗毁性的方法

SDH传输网络亮点之一即是其拥有众多的容错方法。这些方法可以在某些网络元素（可能是通信链路、多路复用器端口或卡，或者多路复用器整体）失效后使网络迅速恢复（在几毫秒内）。

SDH容错机制也有一个著明的通用术语——**自动保护交换**（automatic protection swiching），其反映出在主元素失效的情况下，切换到一个保留路径或者一个保留多路复用器元素。支持此种机制的网络在SDH标准中称为**自愈网络**（self-healing network）。

在SDH网络中，使用如下三种保护方法：

- 1+1保护——即有一个冗余元素执行与主元素相同的工作。例如，当使用1+1方法保护支路卡时，流量同时通过主卡（工作中）和备用卡（冗余）。
- 1∶1保护——即在正常操作模式下，冗余元素并不实施主要元素的功能，而是在主元素失效的情况下接管主元素的工作。
- 1∶*N*保护——为*N*个工作元素（需要保护的单元）分配一个冗余元素。在其中一个主元素失效的情况下，冗余元素接管并开始实施失效元素的功能。如果这种情况发生，那么直到失效元素被替换后其他元素才有保护。

根据冗余元素保护的元素类型，在SDH设备和网络中的自动保护交换分为五种主要类型。

设备保护交换（equipment protection switching, EPS）保护SDH设备的单元和元素，它用于对象关重要的多路复用器元素，如处理器单元、交换单元（交叉连接）、电源器件和同步信号输入单元等的保护。因此，EPS以1+1或1∶1的方式运行。

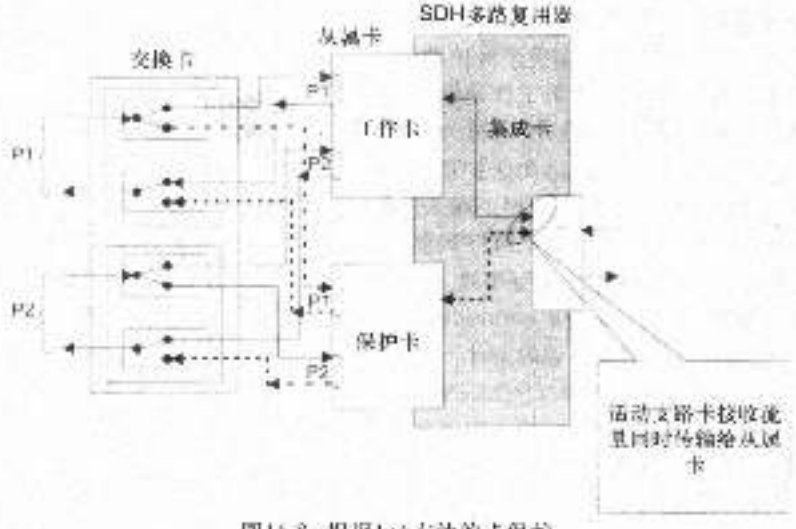


图11-8 根据1+1方法的卡保护

卡保护（card protection）保护多路复用器的聚合卡和从属卡。它允许多路复用器在其中一个聚合卡或者支路卡失效的条件下自动继续运行。在卡保护中，所有的三种方法——1+1、1∶1和1∶*N*——都被采用。采用1+1方法的保护确保传输服务的连续性，因为在卡失效的前提下用户流量并不会间断。图11-8提供了根据1+1方法支持两个支路端口卡保护的多路复用器实例。其中一个支路卡为主（工作）卡，另一个为冗余（保护）卡。以此种方式连接的卡对它们的操作模式用专用命令对多路复用器进行配置。当两个支路卡同时使用时，它们并行处理流量。

对于支路卡间的流量交换，使用从交换卡。从每个端口来（添加）的流量提供给交换卡的输入桥，用于分流流量并将流量提供给支路的适当端口的输入端。聚合卡同时从支路卡接收STM-N信号，并从活动卡选取信号。聚合卡的输出（去往）流量也同时由两个支路卡处理，但只有来自活动卡的流量由交换卡提供给输出。

当主卡失效（或者其他需要切换至保护卡的事件发生时，诸如信号退化、信号错误或者卡移除），由多路复用器控制单元的命令控制的聚合卡，切换到由支路保护卡接收信号。交换卡同时开始传输从保护卡丢弃流量的信号到输出端。

这一方法确保通过保护卡自动保护所有的连接。当指定卡保护时，工作卡连接的配置与保护卡相同。

多路复用段保护（multiplex section protection, MSP）确保多路复用段（例如，相邻SDH多路复用器间的网络段）受到保护。比起卡保护，其更有选择性。保护对象是两个多路复用器间的段，包括两个端口和通信链路（可能包括再生器而没有多路复用器）。通常使用1+1保护方法。当使用这种方法时，为工作链路配置（由电缆连接的端口的高位对）保护链路（端口的低位对），（图11-9a所示）。当建立MSP时，需要在工作端口和保护端口之间通过指定相互关系来配置每个多路复用器。在初始状态时，所有流量都经过工作链路和保护链路传输。

MSP可以是单向或者双向的。当使用单向保护（图11-9所示）时，只有其中一个多路复用器——失效链路的接收端——决定怎样切换到保护链路。检测到失效后（端口失效、信号错误、信号退化等），该多路复用器通过切换至保护链路接收信号。在这种情况下，发送和接收通过不同的端口完成（图11-9b）。

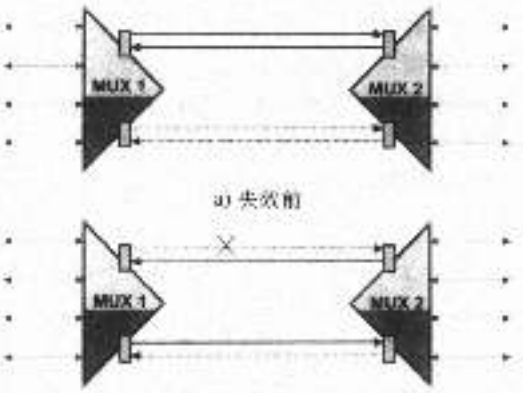


图11-9　多路复用段的保护（MSP）

当使用双向MSP时，如果工作链路在任意方向上失效，那么多路复用器完全切换至保护端口。为了通知发送端多路复用器（使用工作链路）切换至保护链路，接收端多路复用器使用称为K-字节的协议。这一协议，把工作链路和保护链路的状态以及失效的详细信息，插入到STM-N帧头的2个字节中（MSOH的K1和K2字节）。MSP机制确保所有连接的保护都穿过被保护多路复用段。根据标准的要求，交换时间必须低于50毫秒。

子网络连接保护（subnetwork connection protection, SNC-P）为专用虚容器穿越网络提供路径（连接）。它确保在主路径失效的条件下，特定用户的连接切换至备用路径。SNC-P对象把支路流量放入到特殊类型的虚容器中（例如，VC-12、VC-3或者VC-4），使用1+1保护方法。

SNC-P配置两个多路复用器——输入多路复用器，分流放入虚容器的支路流量；输出多路复用器，在这里流量的两条备用路径会合。SNC-P实施的实例如图11-10所示。在ADM1多路复用器中，为T2支路端口的VC-4指定了两条连接：一个连接至A1聚合端口4个VC-4中的一个，另一个连接至A2聚合端口4个VC4中的一个。这些连接中的一个配置为工作连接，一个连接被设置为工作连接；这些

连接中的一个配置为工作连接，另一个成为保护连接。在正常操作模式下，通过两个连接传输流量，传输多路复用（对于这两个连接）以通常的方式配置。在ADM5输出多路复用器中，支路T3端口的VC-4也连接到容器——A1聚合端口的容器和A2聚合端口的容器。从到达T3端口的两个流中选择高质量的流。如果两个流的质量正常或者相等，则选择来自配置工作集成端口的信号。

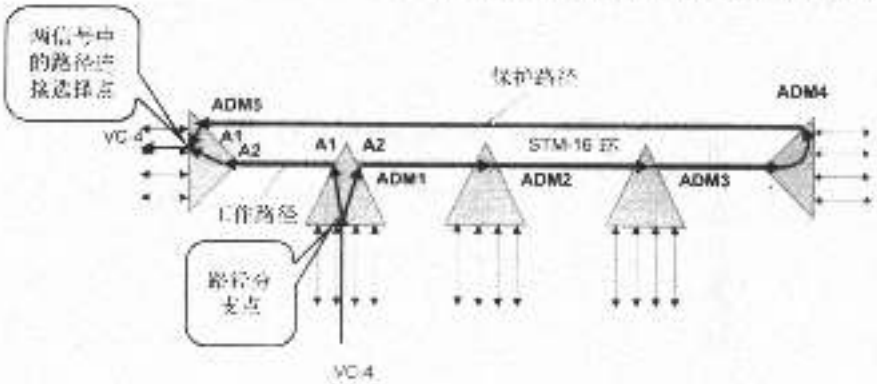


图11-10 SNC-P

SNC-P工作在拥有备用路径拓扑结构（即在环和网状拓扑中）的SDH网络中。

多路复用段共享保护环（multiplex section shared protection ring, MS-SPRing）是在环拓扑中共享用户连接的路径保护，在某些情况下，它确保对环中流量更有效的保护。虽然SNC-P适合于SDH网络的环拓扑，但在某些情况下，它的实现降低了环的有效带宽，因为每个连接沿着环占用双倍带宽。例如，在STM-16环中，根据SNC-P方法，只可能有16条VC-4连接受到保护（图11-11）。

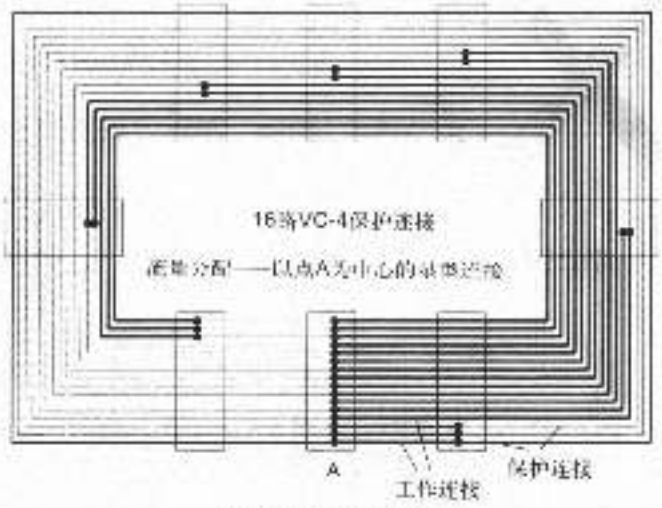


图11-11 环中的SNC-P

MS-SPRing允许更有效地使用环的带宽，因为对于每个连接，都没有事先保留带宽。相反，保留一半的环带宽，这些保护带宽在需要的时候动态地分配给连接（例如，在检测出链路和多路复用器失效后）。使用MS-SPRing对的空闲带宽的级别依赖于流量分布。

如果所有流量都到达同一个多路复用器（即"星型"分布发生），与SNC-P相比，MS-SPRing在经济性方面表现一般。该情形的一个实例如图11-12a所示。在图中，A多路复用器是所有流量集中的中心，且在环中，使用与图11-11所示SNC-P实例相同的16条保护连接。为了保护连接，要保

留了STM-16聚合流的16个虚拟容器中的8个。

在发生故障和失效的情况下，例如链路错误（图11-12b），连接断开的多路复用器流量保存在相反的方向中。这通过使用受影响连接的虚容器所连接的聚合端口的保护虚容器实现的。没有受到故障影响的所有连接在正常模式下继续工作，而不使用保护容器。使用MS-SPRing方法切换到保护连接所需时间为50毫秒。当使用混合流量分布时，MS-SPRing的带宽节省可能更为明显。

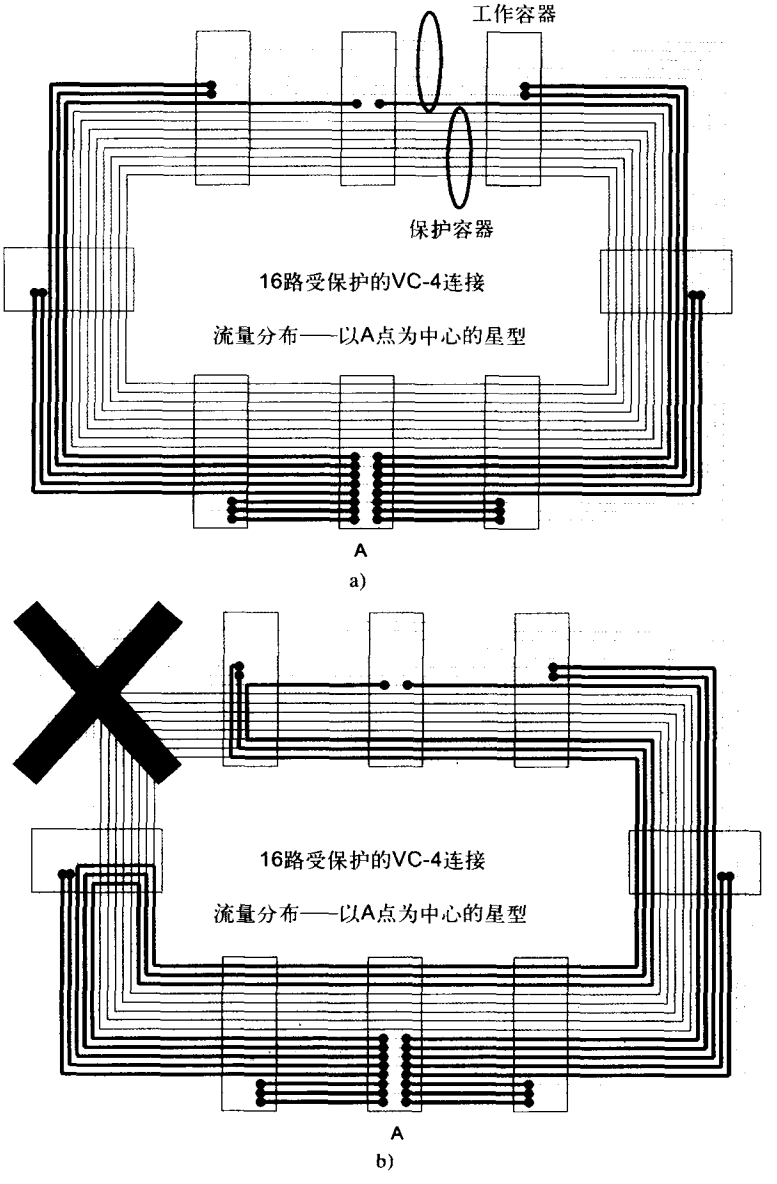


图11-12　MS-SPRing——环共享保护

11.4　DWDM网络

密集波分复用（dense wave division multiplexing, DWDM）技术用于创建新一代的光纤主干网，以数吉字节或者钛字节运行。如此革命性的性能逾越由多路复用方法保证，且与SDH网络中使

用的多路复用方法有本质的差别。在DWDM网络中，光纤电缆中的信息由多个光波同时传输——兰布达（*lambdas*，采用物理中使用的波长的传统定义——λ）。

DWDM网络根据电路交换原理操作，且每个光波是独立的光谱信道。每个光波都承载自己的信息。

DWDM设备并不是直接参与解决每个波长上的数据传输问题——即，选择信息编码和传输协议。其主要功能为多路复用（*multiplexing*）和解多路复用（*demultiplexing*）操作，也就是，在相同信号束中组合不同的波长和从聚合信号中分离出每个光谱信道的功能。最先进的DWDM设备也可以切换光波。

说明 DWDM技术是创新性的，不仅仅因为其通过光纤提高了数据传输速率数十倍；这项技术也开创了多路复用和交换技术的新纪元，因为它不需要将光信号转换成电子形式就可实施这些操作而不需要将光信号转换成电子形式。所有其他类型的在光纤上使用光信号传输信息的技术，诸如SDH或者千兆以太网，都需要将光信号转化成电子信号后，才可以实施多路复用和交换操作。

第一个使用DWDM技术的应用是长距离主干网，用于试图连接两个SDH网络的长距离主干网。采用最简单的点对点拓扑，DWDM设备实现波交换是绰绰有余的。但是，当所采用的技术和DWDM网络的拓扑变得越来越复杂时，这一功能就十分必要。

11.4.1 运行原理

如今，DWDM技术允许一个光纤中使用一个32或者更多种不同波长，在透明窗口为1 550nm的条件下进行传输，每个光波可以以高达10Gb/s的速率承载数据（当使用STM协议或者10G以太网用于在任意波长传输信息时）。此过程中的研究工作旨在把每个波长的信息率增加到40Gb/s和80Gb/s之间。

DWDM的前身是**波分复用**（wave division multiplexing, WDM）技术，其在1 310nm和1 550nm的透明窗口上使用4条光谱信道，载波范围从800GHz到400GHz。（因为没有标准的WDM分类，所以也会遇到其他特性的WDM系统。）

DWDM多路复用成为密集型是由于使用了比WDM明显小的波长间距。如今，ITU-T G.692建议定义两频率（波长）栅格（即以特定值彼此分离的频率集）：

- 频率栅格在相邻信道的频率间隔为100GHz（Δλ = 0.8nm），据此使用从1 528.77nm（196.1THz）到1 560.61nm（192.1THz）范围的41条光波。
- 频率栅格使用在50GHz（Δλ = 0.4nm）的间隔空间，允许在相同范围内使用81种波长传输。

一些公司也生产称为**高密WDM**（high-dense WDM）的设备，它能够在25GHz频率步进下运行（此时，大多是实验设备而不是批量生产）。

实现步进为50GHz和25GHz的频率栅格对DWDM设备有严格的要求，尤其是当每个光波以10Gb/s或者更高（STM-64、10GE或者STM-256）速率调制时。再次值得强调的是，DWDM技术（与WDM相似）不是直接涉及每个波长上承载的信息；这一难题由高层技术解决，它根据需要分配的波长，传输模拟或者离散信息。这些可能是SDH或者10吉以太网技术。

理论上，相邻波长之间50GHz甚至25GHz的间隙允许以10Gb/s的速率传输数据。但在这种情况下，要保证频率的高精确性和最小光谱带宽。而且，要降低噪声级别以便将光谱重叠的影响降到最小（图11-13）。

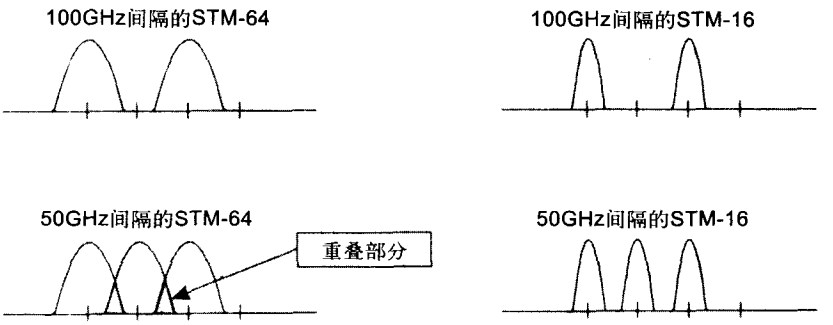


图11-13　不同频率计划和数据传输率条件下的连续波长的光谱重叠

11.4.2　光纤放大器

DWDM技术设备已经作为最顶级通信运营商骨干运行，它的成功实际归因于光纤放大器（*fiber amplifiers*）的出现。这些光设备在1 550nm范围直接放大光信号，因此消除了把这些信号转化为电子形式的中间转换，这些转换由SDH网络中使用的再生器完成。而且，电子信号再生系统的代价昂贵且依赖于协议，因为它们需要理解信号编码的特定方法。光纤放大器“透明地”传输信息，允许不需要升级放大单元即可提高主干速率。

两个光纤放大器单元间的段的长度可以达到150km或者更长，这一点保证了DWDM主干网的经济效率，如果使用1到7个中间光纤放大器，那么多路复用段的长度为600~3 000km。

ITU-T G.692定义了三种放大段（即相邻DWDM多路复用间的段）的类型：

- **长**（long, L）——段由8个光纤通信链路段和7个光纤放大器构成。放大器间的最大距离可以达到80km，可拥有的最大段长度为640km。
- **甚长**（very long, V）——段由不超过5个的光纤通信链路段和4个光纤放大器构成。放大器间的最大距离为120km，且最大段长度为600km。
- **极长**（Ultralong, U）——段不含中间放大器且长度为160km。

被动段的数量和其长度的限制与光放大过程中的光信号退化有关。虽然EDFA光纤放大器存储了信号能量，但它不能完全补偿色散[⊖]和其他非线性作用的影响。因此，为了构建长距离的主干网，需要在放大器间安装DWDM多路复用器。这些DWDM多路复用器通过将光信号转化成电子形式再转换回光形式重新产生信号。为了减少非线性作用，DWDM系统也隐含了信号能量加以限制。

光纤放大器不仅用于增加多路复用器间的距离，而且也在多路复用器内部使用。虽然多路复用和交叉连接使用专用的光工具实现，并没有将信号转化成电子形式，但信号在被动光转化过程中会有能量损失，且它们必须在送入线路之前进行放大。

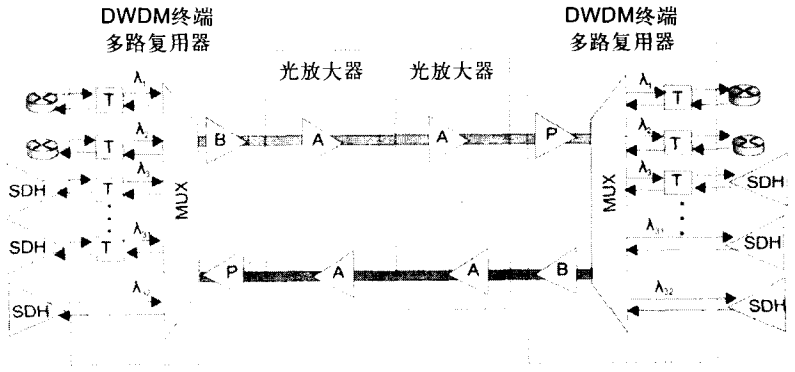
光纤放大器领域的新研究使得出现运行在称为L范围（第四透明窗口）即从1 570nm到1 605nm上的放大器。这一范围的使用，以及50GHz波长到25GHz间空间的减少，使得允许同步传输的波长增加到80或者更高。这意味着确保通过一条光纤在单向上流量传输率达到800Gb/s——1.6Tb/s成为可能。

DWDM的成功导致了另一种前沿技术领域的出现——**全光网络**（all-optical network）。在这种网络中，所有涉及到多路复用/解多路复用、添加/丢弃和交叉交换/路由用户信息而不用将光信号转化为电子形式即可实施。消除将信号转化为电子形式显著降低了网络开销。然而，如今的光技术水平是不足以创建大规模全光缆网络的。因此，该网络的实际应用局限于全光段，段之间的信号依然为电子形式重发。

⊖　色散发生归因于具有不同长度波长的传播速率的不同。因此光纤接收端的信号被“污浊”。

11.4.3 典型的拓扑

从年代顺序上讲，DWDM（与SDH相同）的第一个应用领域标志着拥有点对点拓扑、极长距离和高速主干的建立（图11-14）。



计算机网络设备（路由器，交换机）

图11-14 基于DWDM终端多路复用器的极长距离点对点主干网

为组建这样的主干网，在终端点上安装终端DWDM多路复用器是已经足够了。在传输点，如果终端点之间的距离超过了距离限制，那么必须安装光纤放大器。

终端多路复用器包括：

- 多路复用/解多路复用单元
- 输入放大器，或辅助放大器（B）
- 前端放大器（P）
- 异频收发机集合（T）

异频收发机将波长不符合多路复用器频率栅格的源输入信号转化为所要求长度的波。当连接到DWDM网络的设备能够产生出DWDM多路复用器所支持波长的信号时（根据ITU-T G.692频率栅格或者特定生产商频率栅格）时，就不使用异频收发机。在这种情况下，说连接到DWDM网络的设备拥有一个有色（*colored*）接口。

在前面提到的方法中，网络用户间的双工交换的代价为通过两条纤维的整个光波集合只能单向传输。当节点间只有一条光缆用作通信时，产生DWDM网络操作的另一个变化形式。这种情况下使用光纤保证信号以双工模式双向传输——频率中的一半光波以一个方向传输，另一半光波则用于相反方向的数据传输。

中间连接网络（network with intermediate connection）。其中的传输节点实施添加/丢弃多路复用器功能，它是前一种拓扑的自然改进（图11-15）。

光添加/丢弃多路复用器（optical add/drop multiplexer, OADM）。OADM可以从聚合光谱信号中分离（丢弃）出特定长度的光波，并可以添加同一波长的信号，所以传输信号的频谱不会改变，并且给连接到传输多路复用器的用户建立连接。OADM可以通过光方法或者将信号转化为电子形式来添加和丢弃信号。通常，全光（无源）添加/丢弃多路复用器可以丢弃少量光波。这是因为每个丢弃操作都需要将光信号通过光过滤器，这增加了衰减。如果多路复用器使用电子信号重发，那么可以在可用波长集合限制内丢弃任何数量的光波，因为传输光信号已经在先前完全解多路复用了。

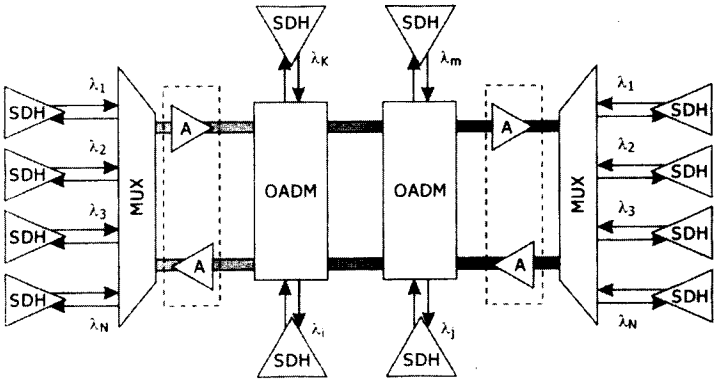


图11-15　传输节点上拥有添加/丢弃多路复用器的DWDM网络

环（Ring）。由于保留路径的存在环拓扑（图11-16）确保DWDM网络的生存。在DWDM中使用的流量保护方法与SDH方法相似（虽然在DWDM中它们还没有标准化）。为了保护一个特殊的连接，它的端点间建立两条路径——主路径和保护路径。端点的多路复用器比较两个信号并选择较好质量的那个信号（或者默认信号）。

网状拓扑（mesh topology）。随着DWDM网络的革新，（图11-17）在设计中，越来越频繁地使用网状拓扑（图11-17），因为与其他拓扑相比它能确保更高的兼容性、性能和容错。然而，为了实现网眼拓扑，有必要采用**光交叉连接**（optical cross-connect, OXC）。OXC不仅同添加/丢弃多路复用器一样从聚合传输信号中添加/丢弃光波，而且支持以不同波长的波传输光信号间的任意交换。

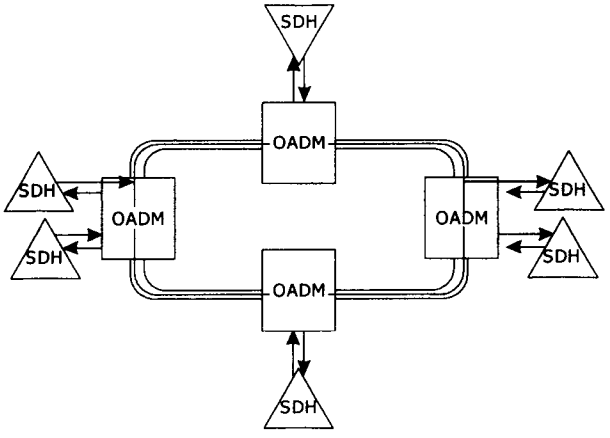


图11-16　DWDM多路复用器的环

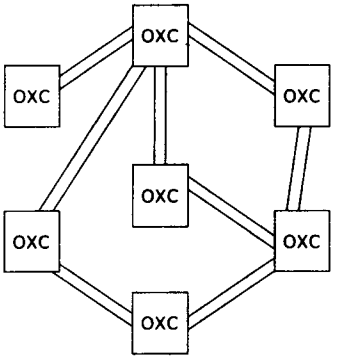


图11-17　DWDM网络的网状拓扑

11.4.4　光添加/丢弃多路复用器

光多路复用器在公共聚合信号内多路复用多个波长，并且从聚合信号中丢弃不同长度的光波。

为了丢弃光波，多路复用器可以使用不同的光机制。支持相对少量波长种类，通常从6到32的光多路复用器，使用**薄膜过滤器**（thin-film filter）。它们包含有多层涂层构成的金属板。实际上，光纤末端倾斜30到45度，且覆盖有作为薄膜过滤器的多层涂层。对于具有较大数量波长的系统，需要其他的过滤和多路复用方法。

在DWDM多路复用中，使用**全衍射相位光栅**（diffraction phase grating），或者**排列波导光栅**（arrayed waveguide grating, AWG）。栅板的功能由光波导或者光纤实现。到达的多路复用信

号提供给输入端口（图11-18a）。然后，信号穿过棱板波导且分送给代表衍射AWG结构的波导集合。在每个波导中的信号依然是多路复用信号，且每个信道（$\lambda_1, \lambda_2, \cdots, \lambda_n$）仍在所有波导中存在。接着，信号从镜面金属盘反射回来，最后，光束又一次集中到棱板波导。这里，它们被聚焦，并产生干扰。因此，干扰的最大值出现，并分布于空间中。这些强度最大值对应于不同信道。平板波导布局——特别是，输出电极的位置和AWG波导的长度值——计算出来，使得干扰最大值和输出电极吻合。多路复用以相反的方式完成实现。

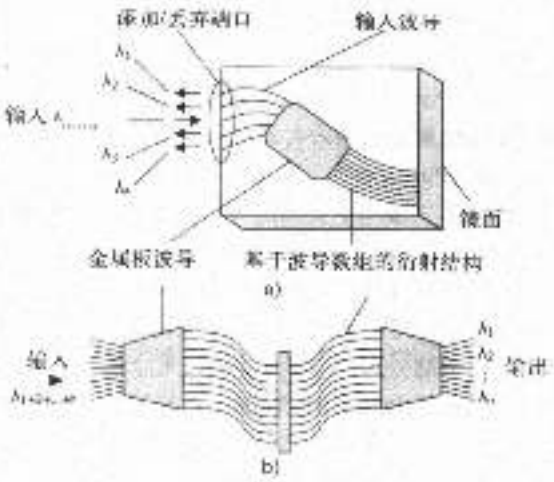


图11-18 使用衍射相位光栅的信号的完全解多路复用

构建多路复用器的另一个方法是基于平板波导（图11-18b）。该设备除了用于调焦和干扰的额外盘外，运行原理与先前的例子相似。

全AWG光栅（也称移相器）成为DWDM多路复用的关键元素。通常，它们采用光信号的完全解多路复用，因为它们能够成功升级且具有在数百个光谱信道的系统上成功操作的潜力。

11.4.5 光交叉连接器

在具有网状拓扑的网络中，必须确保在网络用户间灵活切换光波连接的路由器的能力。这些功能由OXC保证，它允许将任意端口的输入信号的光波指向任何的输出端口（假如这一端口的其他信号没有使用这一光波；否则，必须转化波长）。

OXC分为如下两类：

- 信号转化为电子形式的OXC
- 全光OXC

光电子交叉连接是第一次出现，通常命名为OXC。因此，这一类型的全光设备生产厂商试图为他们的产品使用不同名字，诸如光电子交换机（photonic switche）、光波路由器（wave router）或者兰布达路由器（lambda router）。OXC有原则限制——它们在2.5Gb/s的速率运行时达到最优状态；但从10Gb/s开始，这些设备的型号和它们的能量消耗超过了所有限制。光电子交换不受该种限制。

在光电子交换机中，使用不同的光机制，包括衍射相位光栅和**微电子机械系统**（micro-electro mechanical system, MEMS）。

MEMS是微小移动的镜面集合，直径不足1mm（图11-19）。当源端信号已经被分割成分量光波时，MEMS交换机在解多路复用器之后使用。通过把小镜面旋转一个特定角度，特定波长的源光束传向相应输出光纤，然后，所有的光线被多路复用到聚合输出信号中。

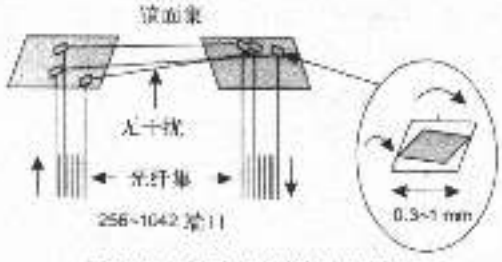


图11-19 用于交叉交换的MEMS

与OXC相比，光电子交换机大约小30倍且节能大约100倍。但这种设备也有缺陷，最重要的是响应慢且对机动敏感。尽管这样，MEMS仍在光电子交换机中被广泛使用。如今，这些设备能够确保为256×256个光谱信道提供交换，且发行能够支持1024×1024或者更高光谱信道的设备近在咫尺。

11.5 案例学习

这个案例学习描述了一个大型能源公司，ABC-Power的传输网络。虽然如同本书中所有提供的案例学习一样，这是一个假定的名字，但这些案例学习是基于真实世界的项目和系统。

ABC-Power公司为覆盖数百平方公里的广泛区域提供电力。公司包括了多个大型产能发电站，和用于提供能量给用户大公司和独立客户的分布网络。

ABC-Power对象，发电站和分布站分散在超过50个城市和定居区域。电源电路为三级层次控制系统：第一级为中央管理节点，第二级是区域控制节点，第三级由发电站和分布站构成。ABC-Power使用不同的工具管理生产和分布能源的过程，包括：

- 遥感和自动控制系统用于控制不同的技术对象（发电站和分布站）。这些系统由传感器构成，传感器用于提供关于电力模式状态和控制它们的启动机制的在线信息，实施诸如从分布网络的一部分到另一部分的电力重分配等操作。遥感数据以实时模式在对象间传输。这些数据也提供给中央或者区域办公室的管理人所使用的中心板。
- 专用分派通信。这是一个与电话网络相似的声音通信系统，带有多种辅助功能，以协同方式帮助调度解决出现的问题。
- 基于PBX的私人电话网络。这一网络补充了调度通信系统的能力，并且拥有至国际电话网络的连接。
- 用于管理公司资源的可定制自动计算机系统。

上述列出的系统构成了分布于所有50个ABC-Power公司网点的子系统。显然，需要高质量的通信网络来确保控制和管理系统的稳定操作，它通过可靠和高速链路连接所有ABC-Power公司的网点。

相当长的一段时间内，ABC-Power从区域通信运营商租赁64Kb/s到2Mb/s的通信链路。这些链路用于连接PBX和局域网的路由器/交换机。遥感和自动控制系统部分使用ABC-Power拥有的铜链路。这些链路沿着电力传输对象的链路安装，这些对象超过了区域通信运营商的服务范围。

ABC-Power业务的进一步发展需要使用最先进的管理方法，包括安装结合调度通信和传统电话功能的新数字PBX电缆，采用强劲的SAP R/3管理系统替代孤立部门管理系统和遥感，自动控制系统的发展。

这些管理工具的现代化需要通信链路的更新换代，包括可靠性的改善和带宽的增加。

对通信链路基础设施升级的可能变体进行分析表明，租赁确保34Mb/s到155Mb/s速率的高速链路在经济上是诱效的。因此，ABC-Power决定使用已存在的电力传输线路网络的优势组建自己的传输网络。这一方法被众多铁路、电力和石油天然气公司所采用。沿着已有铁路或者天然气管道铺设光纤电缆不需要明显的开销，且通常以高投资回报率为特点。

ABC-Power传输网络在两年内建成。在公司现有的所有50个节点内，光

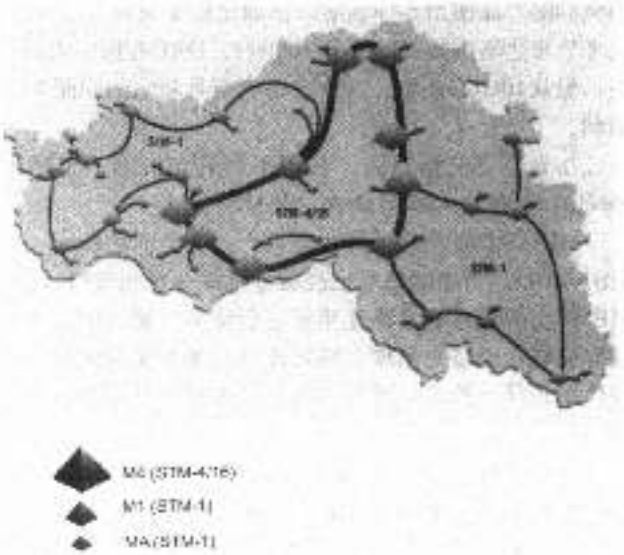


图11-20　ABC-Power公司的SDH网络传输

续连接到SDH多路复用器。

这一传输网络具有网状拓扑，其允许公司采用SDH技术的路径保护方法并保证高可靠性。网络中使用三种多路复用器：M4，M1和MA。M4多路复用器是STM-4级添加/去掉多路复用器——即，它们的聚合端口以STM-4速率（622Mb/s）运行。这些多路复用器形成连接大型区域控制节点和中心控制节点的主干环。M4多路复用器允许用STM-16聚合端口（2.5Gb/s）替代STM-4聚合端口，其运行在DWDM频率的一段光波上。这确保了在不替换设备的前提下，即可实现未来增加主干网速率未来增长的可能，包括把SDH网络连接到DWDM主干网的可能性。

传输访问网络基于M1和MA多路复用器（STM-1 155Mb/s聚合端口），其跨越所有的发电站和小型区域控制节点。访问网络将网状拓扑和树型拓扑相结合，确保只对大多数关键路由的冗余。MA多路复用器依据多个用于连接重叠网络设备的PDH端口而不同。这些设备包括电话，计算机和遥感或控制网络（图11-21）。

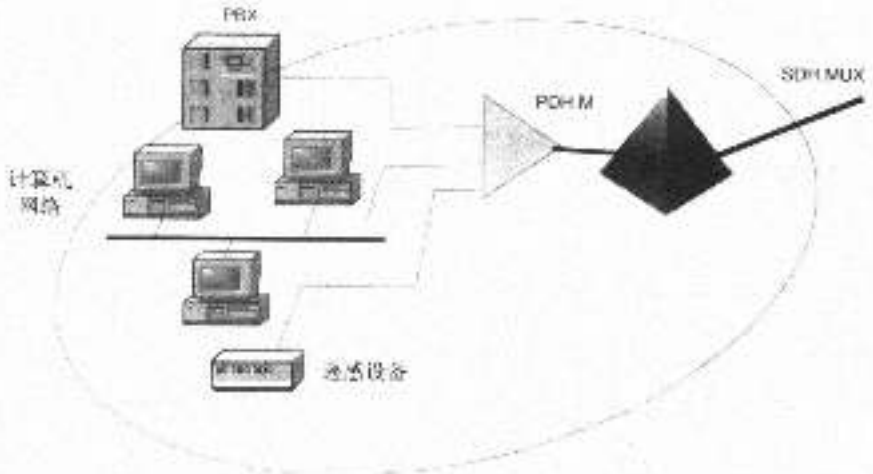


图11-21 设备连接到SDH网络

私人SDH网络的建立允许ABC-Power在公司需要高速通信链路时处于安全状态。公司计划使用剩余链路作为ISP进行商业运作。

小结

- 传输网络用于创建交换基础结构，使用它可以创建任意拓扑的快速、灵活的永久信道。
- 传输网络使用不同类型的电路交换技术，采用频分复用（FDM），时分复用或者波/密波分复用（WDM/DWDM）。
- 在FDM网络中，每个本地环路都分配宽度为4kHz的频率带宽。FDM信道层次结构是，12条本地环路以48kHz构成第一级信道（基础组），5个基础信道以240kHz的带宽连接成第二级信道（超级组），10个二级信道以2.4Mb/s频率带宽组成第三层次（主要组）。
- 数字传输网络（PDH）允许创建输出范围从64Kb/s到140Mb/s的信道，提供给用户四级速率层次。
- PDH网络的缺点为不可能直接分离数据的低速率信道和高速率信道，因为这些信道运行在速率层次的非连接级。
- 把用户流添加到SDH帧中的异步模式由虚容器概念和在虚容器中标记用户数据开始位置的浮动指针系统保证。
- SDH多路复用可以在具有不同拓扑的网络中运行，包括链、环和网状拓扑。有几个特殊类

型的多路复用器在网络中占据着特殊位置：终端多路复用器，丢弃/添加多路复用器和交叉连接（OXC）。

- 在SDH网络中，许多支持容错的机制在特定块、端口或者连接上保护数据流量：EPS、卡保护、MSP、SNC-P或者MS-SPRing。选择最高效的保护方法取决于网络连接的逻辑拓扑。
- 多路复用的WDM/DWDM技术以信号的另一天然属性和速度层次实现了频率多路复用原理。每个WDM/DWDM信道是光波的专用范围，允许以模拟或者数字形式承载数据。同时，信道带宽为25~50~100GHz，保证每秒数吉兆比特速率（当传输离散数据时）。
- 在早期的WDM系统中，使用少量光谱信道，从2到16。在DWDM系统中，单一光纤电缆使用32到160条信道，这确保了在单根光纤上数据率可达到数钛比特每秒。
- 当代光纤放大器允许连接链路的光段无需把信号转化为电子形式就可延伸700~1 000km。
- 为将多个信道从聚合光信号中分离出来，相对廉价的设备通常与光纤放大器结合在长距离网络上组织丢弃/添加多路复用器。
- 为和传统的光网络（SDH、吉兆以太网和10吉兆以太网）交互，DWDM网络使用异频收发机和波长转换器，将输入信号的波长转化为DWDM标准频率设计中的波长。
- 在全光网络中，所有的多路复用和交换操作都在光信号上实施而不将信号转换为电子形式。这简化了网络并降低成本。

复习题

1. FDM传输网络的什么缺陷导致了数字传输网络的产生？
2. T-1名字的意思是：
   A. 由AT&T开发的多路复用设备
   B. 1.544Mb/s速率级
   C. 通信链路的国际标准
   D. 多路复用64Kb/s数字流的方法
3. 传输声音时，在T-1信道中，哪个功能可以指定给每个字节的不重要位？
4. 是否有可能在PDH网络中把DS-0信道从DS-3直接分离出来？
5. 在解决前一问题时，实际采用何种方法？
6. 在实施E-1信道时，采用何种机制替代T-1信道的位劫持？
7. 为什么传输网络可以保证所有流量的服务高质量？
8. 由“伪年代顺序”术语反映的是PDH技术的什么属性？
9. SDH技术怎样补偿支路流的同步缺失？
10. STM-1帧可多路复用E-1信道的最大值是多少？
11. 如果STM-1帧已经包含15条E-1信道，那么它可多路复用多少T-1信道？
12. SDH协议栈的哪一层在设备失效的条件下负责网络配置？
13. SDH再生器间的数据通信信道的最大速率为多少？
14. 为什么STM-1帧使用3指针？
15. 在SDH和PDH技术中使用交叉字节的目的？
16. 1+1和1:1保护方法有何不同？
17. 什么条件下MS-SPRing比SNC-P更有效？
18. FDM和DWDM传输网络的共同特点是什么？
19. DWDM网络是什么类型的网络——模拟还是数字？
20. 在DWDM网络中，使用再生器将光信号转化为电子形式的目的是什么？

21. 当传输多个无源DWDM段时，光信号退化的原因是什么？
22. 在OXC中，交换光信号的原理是什么？

练习题

1. STM-1帧中的VC-4点的负校正频率是多少，如果发送和接收SDH多路复用器的时钟频率相对差为10 ⁻⁹？
2. SDH网络由四个STM-1多路复用器：*A*、*B*、*C*和*D*构成。图11-22显示了多路复用器间的流量分布。所有流为STM-1速率。多路复用器连接至STM-4环。应选择何种保护方法保护所有的连接？

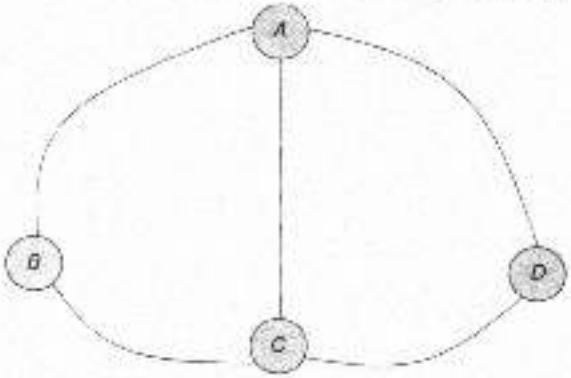


图11-22 流量分布

第三部分 局域网

LAN是当代所有计算机网络的重要组成部分。如果观察某个广域网的体系结构，例如因特网或者更简单的一个大规模的企业网，你会发现所有网络的信息资源都集中在LAN中，广域网只是连接众多LAN的传输器。

LAN的最重要的目的之一是连接计算机。这些计算机在一栋楼里或是在彼此相隔不远的一组楼群中，通过本地服务器提供网络用户信息访问服务。LAN也提供一种简便的方法将计算机分组连接到广域网。广域网在网络间传送信息比在单个计算机间传送要容易。机场和火车站的无线LAN服务就是一个很好的例子：通常，这样的网络不是为了确保临时用户间的信息转发，相反，其主要目的是使用户能够访问因特网。需要强调的是，这些例子中，组建因特网访问是为了整个LAN而不是为某个用户服务。LAN也用于一些电信网络，例如电话和传输网。管理电话接线器的系统和传输网络都是以在LAN为基础的，LAN连接着网络管理者的计算机，确保他们能够访问嵌入在电信网络设备中的控制程序。

LAN技术经历了相当长的发展过程。实际上，20世纪80年代的所有技术都是在物理层使用共享介质（*shared media*）方法连接计算机，这是一种便捷而且成本低的技术。共享介质的基础理论已经在第2章介绍过，本书的这部分我们重新回到这个问题，并从标准和具体的算法层次做更详细的分析。

20世纪90年代中期，LAN开始引入技术的交换版本（*switched version*），摒弃使用共享介质技术从而提高了LAN的性能和可延拓性。交换LAN使用的协议与共享介质LAN是一样的，只是采用全双工的工作方式。交换LAN的另一优势是能够提供多种方法确保服务质量（QoS），当LAN传送实时业务时，这点显得尤其重要，例如IP电话业务。

尽管交换LAN十分流行，但共享介质仍然经常用于新旧技术中。它在有线LAN和无线LAN的小段范围内是有效的，其传送介质原本就是共享的。

LAN使用的介质在不断变化。LAN协议的最大信息传输速率也在增加。2002年，随着10G以太网标准的采用，LAN开始支持速率层次以适应不同的网络传输速率：从10Mb/s到10Gb/s。在这些技术基础上建立MAN和LAN成为可能。

LAN正朝着小型化的方向发展。一种新的网络，个人区域网（*persona area network*）已经出现，这种网络可连接几百米范围内的个人用户电子设备。

现代LAN由一种网络技术所统治，更确切地说，是被一个完整的网络技术系列统治：以太网。本书中自然会重点介绍这个技术系列。第三部分包含下列章节：

- 第12章（以太网）涵盖了经典的基于共享介质的10Mb/s以太网技术。
- 第13章（高速以太网）介绍基于共享介质的高速以太网技术：快速以太网和千兆以太网。
- 第14章（共享介质的LAN）描述基于共享介质的其他LAN技术（令牌环和FDDI）和两种无线网络技术：IEEE 802.11和蓝牙。

最后两章关注交换LAN。

- 第15章（交换LAN基础）涵盖交换LAN运行的基础：LAN交换的工作算法，LAN协议的全双工方式，LAN交换的具体执行特征。
- 第16章（交换LAN的高级特性）介绍这种类型LAN的高级特性，包括以生成树算法为基础的冗余链路，链路聚合和VLAN技术。

第12章 以太网

12.1 引言

以太网（Ethernet）是当今最常用的LAN标准：使以太网协议的网络数量估计有几百万。

以太网这个术语代表多种不同的技术；包含快速以太网、千兆以太网和10G以太网。

从狭义上说，以太网（*Ethernet*）出现在20世纪70年代末，它作为一种传输速率为10Mb/s的数据传输网络标准而成为三个公司的专用标准：数据设备公司（DEC），英特尔公司和美国施乐公司。20世纪80年代初，以太网被IEEE802.3工作组标准化，从那时起，它就成为一个国际标准。以太网是第一个建议使用共享介质进行网络访问的技术。

LAN是分组交换网络，采用时分复用准则，这意味着它们要轮流使用传输介质。介质访问控制（MAC）作为介质共享算法是任何LAN技术中最重要的特性，因为它对技术类型的影响比对信号编码方法和帧格式的影响要显著得多。以太网使用随机访问方法作为介质共享机制：尽管它还很不完善，因为随着网络负载的增加，有效网络带宽急剧减少，但以太网的最大成功在于它的简单。

10Mb以太网的普及也有力地促进了它的发展：1995年采用快速以太网标准，1998年出现千兆以太网标准，2002年达到10G以太网标准。每一个新标准都比前一标准快了十倍，创造了激动人心的速率层次：

10Mb/s—100Mb/s—1 000Mb/s—10 000Mb/s。

这一章，我们具体介绍经典的10M以太网技术，其大多数准则都使用较高的传输速率。

12.2 LAN协议的一般特性

以太网属于LAN整个技术系列中的一种，LAN还包括令牌环、FDDI和100VG-AnyLAN[⊖]。尽管它们有些特殊性质，但这些技术都有共同的目标：组建LAN。因此，从以太网开始学习并了解LAN技术发展的一般特性是很有意义的。

12.2.1 标准拓扑和共享介质

20世纪70年代末，最初LAN的研究人员主要目的是，找到一种简单而又廉价的方法连接同一栋楼里的数百台计算机使它们成为一个计算机网络。这种解决方案必须是廉价的，因为这种网络相连的是已出现并迅速扩散的（花费为每台$10 000~$20 000）低价的微型计算机（相对主机而言）。单个机构中的计算机数量非常小，因此，几百台计算机的上限对任何实际的LAN都足够了。目前，不考虑将LAN连到广域网的问题，因此实际上所有LAN技术都忽略了它。

为了简单化，和减少对软件和硬件的投入，早期LAN的研究人员决定使用**共享数据传输介质**（shared data transmission media）。

Norman Abramson于20世纪70年代初，在夏威夷大学首次使用ALOHA无线电网络测试这种计算机间通信的组织方法。所有发送端使用某个频带进行数据压缩，这个特殊频带的无线电信道就是一个共享介质。ALOHA网络使用随机访问方法，任何节点在任何时候都可以传输分组。如果节点在一段时间后未收到确认信号，它就重传那个分组。共享介质是一个载波频率为400MHz带宽为

⊖ 100VG-AnyLAN技术现在已淘汰，但它用到的介质共享的原始概念仍有理论意义。

40KHz的信道，保证数据传输率为9 600b/s。

不久以后，*Robert Metcalfe*把共享介质的思想用于有线LAN：一个连续段的同轴电缆如同一个公用的无线电介质。所有计算机按线路的设计连接到这个段（图12-1）。因此当一个传送器发送信号时，所有接收端都收到同样的信号。就好像使用无线电波一样。

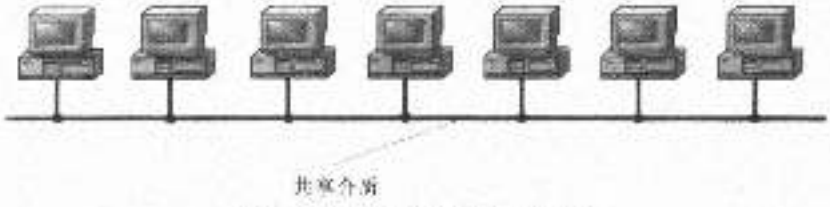


图12-1 基于同轴电缆的共享介质

在令牌环和FDDI中，所有计算机使用的共享介质并不是以太网宣称的那样。这些网络是基于物理环路的拓扑结构，每个节点通过电缆与两个邻近节点相连（图12-2）。然而，这些电缆是共享的，因为任何时候只有一台计算机可以使用环路传送数据。

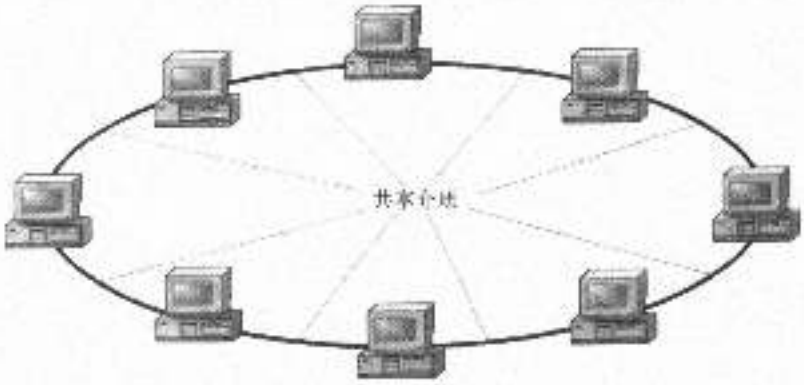


图12-2 环拓扑中的介质共享

简单的**物理连接的标准拓扑**（standard topologies of physical connection）（区别以太网的星型拓扑、令牌环网与FDDI的环形拓扑）程序方便地使用电缆作为共享介质。

使用共享介质简化了网络节点的操作逻辑。因为某个时间只能进行一次数据传输操作，转接节点没必要对帧进行缓存。同样，也没有转接节点（分组交换网的这种特性已在第2章讨论）。因此，复杂的流量管理和拥塞控制过程也取消了。

共享介质的主要缺点是可延伸性差（*poor scalability*）。这种缺点是很严重的，因为介质的带宽分给了网络中的所有节点，不管它们使用何种介质访问方法。虽然如此，使用第7章描述的队列理论得到的结果是可行的：一旦共享介质的利用率超过一定阈值，访问介质的队列开始非线性增加。因此，网络实际上不可用。利用率的阈值取决于使用的访问方法。例如，ALOHA网络中，阈值非常低（大约18%）。以太网中，阈值比较高（大约30%），令牌环和FDDI中，阈值达到60%~70%。

12.2.2 LAN协议栈

LAN技术仅仅执行OSI模型的最低两层功能——物理（*physical*）层和数据链路（*data link*）层（图12-3），因为这两层的功能在标准LAN拓扑结构下发送帧足够了，如总型（总线）、环和树。

但是，这并不意味着LAN连接的计算机不支持数据链路层以上的高层协议。这些协议安装在网络节点中并通过它们执行。但是网络节点的功能与特定的LAN技术没有关联。网络和运输层协议

对LAN节点是必须的，LAN节点要和连接其他LAN的计算机通信，其路径可能包括WAN链路。如果能确保计算机间的交互操作仅限于单个LAN，那么应用层协议可以直接通过数据链路层操作。但是，由于这种受限的通信能力无法满足用户需求，每个连接到LAN的计算机必须支持整个协议栈。因此，一种网络层协议（如IP或者IPX）通过逻辑链路控制层（LLC）执行。另外，不仅是物理层协议和数据链路层协议，整个协议栈都要嵌入在LAN的端节点，以确保应用程序的兼容性。应用程序必须在任何网络环境下正常运行（至少，它们的运行不依赖网络的大小，不管是小的单段LAN还是大规模的路由网络）。

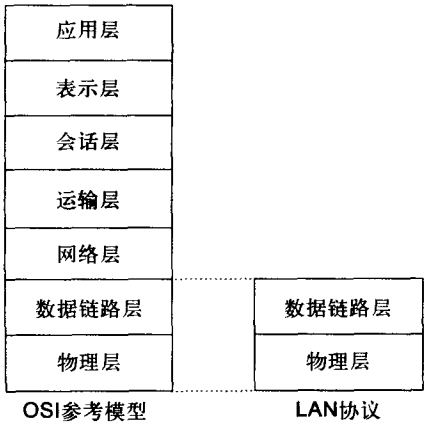


图12-3　LAN协议对应OSI参考模型

LAN中的数据链路层分为两个子层，通常也称为层：

- 逻辑链路控制（LLC）
- 介质访问控制（MAC）

LLC的功能通过操作系统中适当的软件模块执行，MAC的功能则由硬件（网络适配器）和软件（网络适配器驱动程序）执行。

1. MAC层

MAC层的主要功能是：

- 保证访问共享介质
- 使用物理层设备的功能在端节点间传送帧

随机访问（Random access）是访问共享介质的主要算法之一。它的思想是，当一个节点要传送帧时，就使用共享介质而不需要和其他节点协调。

随机访问方法采用分散式（*decentralized*），因为不需要某个特殊节点在网络中充当仲裁者的角色。所以，随机访问方法的特点是发生冲突的可能性大。**冲突**（collision）是几个站点[⊖]试图同时传输帧。冲突导致几个传输器发出的信号重叠，使所有传送的帧中的信息在冲突期间失真。因为LAN使用非常简单的信号编码方法，所以它们不要求所需信号能从信号集中分离出来，与之相反，CDMA就能做到这点。

有许多随即访问算法能够减少冲突发生的可能性从而提高网络性能。例如，一类算法允许一个节点仅在特定的时间间隔内传输帧，通常称为时间槽。这种改进首先由ALOHA提出，此算法的修订版就是著名的时间槽ALOHA。在时间槽ALOHA中，用时间槽的开始时刻同步帧传输（*Synchronizing frame transmission*）使冲突发生的概率减半，确保了在介质利用率达到36%时能进行正常的网络操作。

开始传输前引入载波侦听过程（*carrier sensing procedure*）是对随机访问方法的另一改进。如果节点侦测到介质正忙于传输其他帧时，这些节点就不允许发送帧。尽管没能完全消除冲突，但却减少了冲突发生的概率。

随机算法不能保证某个特定节点在特定时间间隔内访问介质。不管所选的时间值多大，实际等待时间间隔超过它的概率总是大于零。随机访问也不能为不同种类业务量提供有区别的服务质量（QoS）。在任何情况下，所有帧访问介质的概率是相同的。

确定性访问（Deterministic access）是另一种流行的共享介质访问方法。它是由于预先知道访问共享介质的最大等待时间而得名的。

⊖ 本书中，站点和节点具有同样的意义。

确定性访问算法使用两种机制：令牌传递和轮询。

令牌传递（Token passing）通常在分散式（*decentralized*）方法的基础上执行。每台计算机只有当接收到令牌后才能在一个固定的时间周期内使用共享介质，这个时间周期称为令牌持有时间间隔（*token-holding interval*）。计算机可以在这个时间内传送帧，当此时间间隔过去了计算机必须把令牌传给另一台计算机。因此，如果网络中的计算机数量是已知的，那么最大等待时间就等于令牌持有时间间隔乘以网络中计算机的数量。实际等待时间可能更短，因为当某台计算机得到令牌却没有帧要传输时，它就将令牌传出，不需要等到令牌持有时间间隔结束。

令牌从一台计算机传到另一台计算机的次序可由不同的方法定义。令牌环和FDDI中，用链路拓扑定义这种次序。一台计算机在环形网络中有两个邻居：上游（*upstream*）和下游（*downstream*）。它从上游邻居收到令牌并把令牌传给下游邻居。令牌传递算法可以在非环形的网络拓扑中执行。例如，已经过时且不再使用的*Arcnet*，用同轴电缆（如以太网）连接计算机，使用令牌传递访问介质。令牌按预先定义的次序从一台计算机传到另一台计算机，不依赖计算机连接到电缆的位置。

轮询算法（Polling algorithm）通常基于集中式（*centralized*）方法。网络中有个专用节点起着共享介质仲裁者（*arbitrator*）的角色。

仲裁者周期性地询问其他网络节点是否有帧要传输。确认收到的要求后，仲裁者决定哪个节点可以使用共享介质。然后，它通知被选定的节点，这个节点就可以传送帧到共享介质。当帧传送完以后，重复轮询阶段。

轮询算法也可以是基于分散式（*decentralized*）方法。这种情况下，所有节点必须预先互相通知需要使用共享介质传输帧。然后，按照特别的规则，每个独立的节点确定它在帧传输序列中的位置。当轮到这个节点时，它就把帧传输出去。

确定性访问不同于随机访问的是，在网络负载高的情况下，当介质利用率接近1时，算法更有效。另一方面，网络负载轻时，随机算法更有效。因为它允许帧立即传送而不用花费时间决定哪个节点有权访问介质。

确定性介质访问算法的优势在于能区分业务量的优先次序。有了这种能力，就能确保QoS支持。

帧的发送由MAC层执行。这个过程包含多个阶段，与所选的访问方法无关。

- 帧格式化（*Frame formatting*）。这个阶段，帧字段填满从高层获得的信息。这些信息包括源地址和目的地址、用户数据和所发送数据的高层协议代码。当这个帧创建后，MAC层计算它的效验和并放置在对应的字段中。
- 帧传输到介质（*Frame transmission into the medium*）。帧创建后，当节点访问共享介质时，MAC层就把帧传递到物理层，物理层逐字节传送帧的所有字段到传输介质。网络适配器的发送器执行了物理层的功能。发送器把帧的所有字节转换为位序列，使用适当的电或光信号对其编码，然后将它们传送到物理层。信号经过传输介质后，到达和共享介质相连的网络适配器中的接收端。接收端执行逆过程将信号转换为帧字节。
- 帧接收（*Frame reception*）。与共享介质相连的每个网络节点的MAC层检验刚才发送帧的目的地址。如果这个地址与接收端的地址相匹配，那么目的节点的MAC层继续执行，否则就丢弃该帧。更进一步的操作包括对帧的CRC校验。如果收到的帧的CRC正确，MAC就把它传给协议栈的高层。如果帧的CRC无效，那么意味着信息在传输过程中被破坏，这个帧必须丢弃。

基于以上描述，可以清楚知道以太网执行**半双工的数据报传输**（half-duplex datagram transmission）模式。

2. LLC层

LLC层执行下列两种功能：

- 建立与它相邻的网络层的接口。
- 按预定的可靠性级别确保帧传输。

LLC层的接口功能（*interface function*）包括在MAC层和网络层之间传送用户和控制数据。当数据从顶层到底层（*from top to bottom*）传送时，LLC层收到一个包含用户数据的分组（如，IP或IPX分组）。除了这个分组外，网络层还用适当的LAN格式传送目的节点地址。这个地址用于在LAN内传送分组。按照TCP/IP栈，这种地址称为硬件地址。LLC层将来自网络层的数据向下传送到MAC层执行。除此之外，LLC层也执行多路复用（*multiplexing*），因为经过几个网络层协议接收到的数据传送到单个MAC层协议。

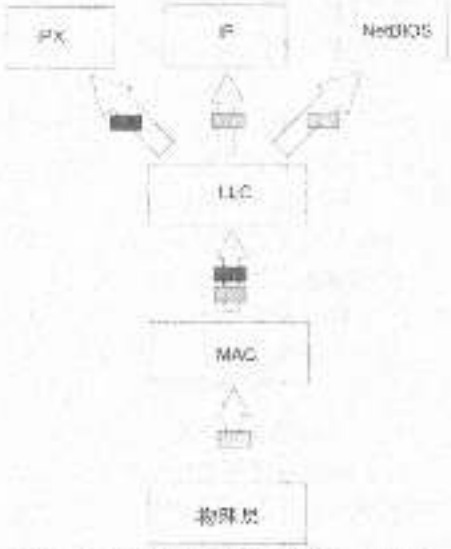


图12-4　LLC协议所实现的解多路复用帧

当从底层向顶层（*frame bottom to top*）传送数据时，LLC层收到来自MAC层的用户数据（如，收到来自网络的网络层分组）。然后，LLC层必须实现附加的接口功能，也就是说，由它决定采用哪种网络协议发送收到的数据。这是一个解多路复用（*demultiplexing*）任务，来自MAC层的集合数据流必须按计算机支持的网络协议的数量分成若干个子流（图12-4）。

多路复用和解多路复用任务不仅是LLC协议的特性，而且也是任何其他上层协议的特性，在此基础上那些高层协议才能执行。为了数据解多路复用，LLC协议在帧的头部使用特殊字段（图12-5），**目的服务访问点**（destionation service access point，DSAP）字段用于存储数据（如，数据字段的内容）到达的协议代码。相应地，**源服务访问点**（source service access point，SSAP）字段用于指定所发送数据的协议代码。使用两个字段实现解多路复用并不是典型的；协议通常只能处理单个的字段。例如，IP通常把它的分组发送给IP，IPX发送给IPX。当一个高层协议支持多个操作模式时两个字段都有效，这样发送节点使用不同的DSAP和SSAP值通知接收节点转到另一种操作模式。NetBEUI协议经常使用LLC的这个特性。

目的服务访问点（DSAP）	源服务访问点（SSAP）	控制字段	数据

图12-5　LLC帧格式

考虑LLC层的第二个功能：确保可靠的帧传输（*ensuring reliable frame delivery*）。LLC协议支持多种操作模式，它们不同之处是当帧丢失或出错时有没有恢复帧的过程。这些操作模式提供的运输服务质量是不同的。LLC层直接与网络层相连，接收网络层的命令，执行数据链路层有特定质量要求的操作。

说明　很明显，LLC的功能是确保数据在LAN中的可靠传输，与OSI模型的运输层功能相似。尽管LLC层不直接涉及网络节点间的帧传输（如运输层所定义），帧访问共享介质后，帧的传输就交给了MAC层。MAC层在数据报模式下执行传输，这意味着不建立逻辑连接并且不能恢复丢失或损坏的帧。如果高层协议要求可靠的运输服务，必须向LLC请求。LLC建立与目的节点的连接并组织帧的发送。

LLC层向高层提供三种类型的运输服务。

- LLC1——无需发送确认的无连接服务。LLC1提供给用户最小代价的数据传输工具。在这种情况下，LLC支持数据报操作模式，与MAC相似，因此整个LAN技术使用数据报模式操作。出错后恢复数据和高层协议执行数据命令时使用这个进程，因此不须在LLC层备份数据。
- LLC2——具有出错和丢失帧恢复功能的面向连接的服务。LLC2使用户能够在开始传输帧之前建立逻辑连接。如必要，它允许进程执行恢复损坏或丢失的数据块（*restoring corrupted or lost data blocks*），并要求这些块在已建立连接的框架内执行。LLC2使用滑动窗口算法实现此目标。
- LLC3——带发送确认的无连接服务。在某些情况下，建立逻辑链路的时间花费是不希望发生的，但发送数据正确接收的确认信号是必须的。控制工业设备的实时管理系统是个很好的例子。LLC3正是为这种情况而设计的。这种服务是LLC1和LLC2的折衷，因为它无需确认连接但要确认数据接收。

LLC操作模式的选择取决于高层协议的要求。LLC传输服务的信息通过层间接口传到LLC层，LLC层带有硬件地址和使用层间服务接口的用户数据。

例如，在TCP/IP协议栈中，确保可靠数据传送的任务由TCP执行，LLC总是在LLC1模式下操作，执行简单的从帧中恢复分组的操作并将它发送给一个高层协议栈。

在所有用到的协议中，仅有基于NetBIOS/NetBEUI的Microsoft/IBM协议栈使用LLC2模式。NetBIOS/NetBEUI协议本身必须执行这种模式以保证能够恢复丢失和出错的帧。所有这些操作由LLC2层承担。如果NetBIOS/NetBEUI协议执行数据报模式，就使用LLC1。

12.2.3 IEEE 802.x标准的结构

1980年，IEEE成立了IEEE 802委员会：目标是LAN技术的标准化，IEEE 802.x的标准体系就是它努力的结果。IEEE 802.x包含了关于LAN底层设计和执行的建议。这些标准的开发是基于一些流行的专用网络标准，例如以太网、Arcnet和令牌环。

IEEE 802委员会的工作成果为国际标准系列如众所周知的ISO8802-1....x奠定了基础。IEEE 802委员会是当今发展LAN技术标准的主要委员会。

还有其他组织也涉及LAN协议的标准化。例如，ANSI为光纤网络开发了FDDI标准，确保数据传输率100Mb/s。它是第一个达到这个速率的LAN协议，比传统以太网速率快十倍。

IEEE 802.2的结构如图12-6所示。

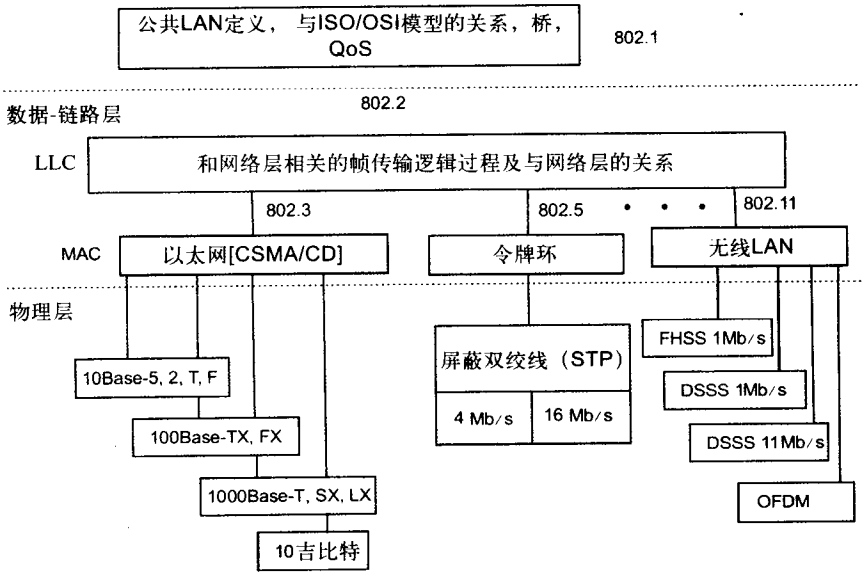


图12-6 IEEE 802.x标准的结构

图12-6中所有技术的MAC层之上是LLC层，这很常见并且对任何特定LAN技术都是独立的。IEEE 802.2工作组（WG）掌管LLC标准。甚至IEEE 802.2委员会框架之外的标准化技术（如ANSI的FDDI协议）都趋向使用802.2标准定义的LLC协议。

标准里的每种技术分两部分描述：MAC层和物理层。如图12-6所示，实际的技术包含多种不同的物理层协议的变体，每个协议对应一个MAC层协议。限于篇幅，图12-6只列出了以太网和令牌环。尽管如此，所有描述对其他技术同样成立，如Arcnet、FDDI、快速以太网、千兆以太网和10G以太网。

IEEE 802.1工作组开发的标准占有特殊重要的位置，因为它们对所有的技术都通用。例如，802.1子委员会提供了LAN的通用定义及其特性和IEEE 802模型的三层与OSI模型的关系。实际上，最重要的802.1标准描绘了不同技术之间的相互关系并为在典型网络拓扑上建立更加复杂的网络奠定了基础。这组标准在通常称为**网络互联标准**（internetworking standar）。它包括了重要的标准，如802.1D，它描述了透明网桥或交换机的操作逻辑；802.1H标准，描述了交换式网桥的操作，能够不用路由器将以太网连接到FDDI或令牌环网；等等。IEEE 802.1工作组开发的标准不断增长。例如，最近它增加了两个重要的标准：802.1Q，定义了在交换网络建立VLAN的方法；和802.1p，描绘了数据链路层流量优先性的方法（如，确保支持QoS机制）。

802.3，802.4，802.5和802.12工作组标准描述了LAN标准；它们是这些标准基础的专有技术不断发展的结果。例如，802.3标准的基础是施乐公司于1975年开发并运行的实验用的以太网。1980年，美国数据设备公司、英特尔和施乐公司（即DIX联盟）开发并出版了以太网标准的第二版本，目的是建立以同轴电缆为基础的LAN。以太网的这个版本就是闻名的以太网DIX或以太网Ⅱ。以太网DIX标准又反过来作为IEEE 802.3标准发展的基础，后者在许多方面与前者相似。802.4标准作为Datapoint公司开发的Arcnet技术的换代产品而出现。802.5标准适应IBM公司设计的令牌环技术。

802.11工作组涉及无线LAN的开发，该无线LAN使用的介质访问方法与以太网中用到的方法近似。因此，802.11标准成为了无线电以太网标准（尽管以太网这个词并未出现在802.11标准中）。

最初的专用技术和它们的改进版本——802.x标准——共存了很长时间。例如，Arcnet并未完全遵从802.4标准组建（现在这样做太晚了，因为Arcnet设备的产品在1993年就停止使用）。唯一的例外是以太网。最新的专用以太网标准是以太网DIX版本Ⅱ。从那时起，没有制造商试图继续开发专用以太网。以太网体系中所有创新的出现都是IEEE802.3委员会采用开放式标准的结果。

后来的标准是由感兴趣公司的团队开发的，然后送到专门的IEEE 802工作组审批。这种技术的例子如快速以太网、100VG-AnyLAN和千兆以太网。首先，有兴趣公司的团队组成一个联盟，其他公司也可以加入到标准的开发过程中。因此，标准的开发过程就自然地开放了。

12.3 CSMA/CD

基于以太网的网络使用一种特殊方法，即（CSMA/CD），访问数据传输介质。

12.3.1 MAC地址

MAC层使用唯一6字节地址确保访问介质和帧传输。这个地址即是IEEE 802.3标准定义的**MAC地址**（MAC address）。MAC地址通常使用六对十六进制数表示，之间用横线或冒号分开，例如：11-A0-17-3D-BC-01。每个网络适配器至少有一个MAC地址。

除单独的接口外，一个MAC地址能定义一组接口甚至整个网络接口。目的地址中最高有效字节的第一（最低有效）位说明这是单地址或是组地址。如果这位是0，就是**单播**（unicast）（单播）地址确定单个网络接口。如果这位是1，则是一个**多播**（multicast）（组）地址。多播地址仅包含组成员的配置接口（管理员手工配置或高层协议命令自动配置），组成员的数量在多播地址中确定。

如果一个组包含一个网络接口，就等效于单播MAC地址，有另一个和它相关的地址，称为多播地址。如果一个多播地址仅由1构成（即，十六进制表示0xFFFFFFFFFFFF），它定义所有的网络节点，称为**广播地址**（broadcast address）。

地址中最高有效字节的第二位定义了地址分配方法：**集中**（centralized）或**本地**（local）。如果这位是0（标准以太网设备几乎都这么做），那么地址就按IEEE 802规则集中式分配。

说明 IEEE以太网标准中，字节的最低有效位在字段的最左端，最高有效位在最右端。字节中这种非标准的位顺序对应了以太网发送端将它们传送到通信线路的次序（最低有效位先发送）。其他组织的标准，如RFC IETF、ITU-T和ISO，使用传统字节表示法，最低有效位在最右端而最高有效位在最左端。同时，传统字节顺序被保留。因此，当读到这些组织发行的标准，通过操作系统或协议分析器在屏幕上显示的数据进行解释时，每一个字节的值必须反转，这样才能得到按IEEE文件规定的每一位的正确含义。例如，IEEE符号表示的多播地址1000 0000 0000 0000 1010 0111 1111 0000 0000 0000 0000 0000（或用十六进制表示80-00-A7-F0-00-00），很有可能通过协议分析器显示为传统形式（如，01-00-5E-0F-00-00）。

IEEE委员会分配**机构唯一标识符**（organizationally unique identifier，OUI）给每个设备制造商。每个制造商把分配的标识符放在地址的前三个字节（如，0x0020AF对应3COM公司，0x00000C表示思科公司）。设备制造商负责确保地址最低三位的唯一性。这24位分配给制造商用于其产品的接口地址，一个机构标识符大概可容纳一千六百万个接口。集中式分配地址的唯一性被广泛用于主要LAN技术，包括以太网、令牌环和FDDI。网络管理员分配本地地址并确保其唯一性。

12.3.2 介质访问和数据传输

为了简单，当考虑用CSMA/CD算法访问共享介质时，假设每个节点（站点）仅有一个网络接口。

任何一台计算机开始向共享介质传送数据时，共享介质网络中的所有计算机都能立即接收到数据（考虑到了通过物理介质传输信号的延时）。所有工作站访问介质的这种操作模式称为**多路访问**（multiple access，MA）模式。

为了得到传送帧的机会，发送接口必须确认共享介质处于空闲状态。这可以通过侦听信号的主谐波频率实现，这个频率称为**载波频率**（carrier frequency）。相应地，这种方法称为**载波侦听**（carrier sense，CS）。载波频率没有出现标志着可以使用介质。假设用曼彻斯特码，这是所有不同类型的10Mb/s以太网都采用的编码，依照传送的1和0序列载波频率分别为5~10MHz。

如果介质是空闲的，那么节点可以开始传输帧。如图12-7所示，节点1检测到介质是空闲的并开始传输帧。在基于同轴电缆的经典以太网中，节点1的发送端发出的信号向两个方向传播以便所有的网络节点都能收到它们。数据帧都伴随着一个**前同步码**（preamble），每个前同步码内容是前7个字节均为10101010，和第8个字节为10101011。最后一个字节称为**帧起始字节**（stat of frame byte）。前同步码对于发送端和接收端之间的位和字节同步是必须的。以上两个值出现，且第二个值紧随第一个值，表明前同步码已经结束，接下来开始传送帧的位。

所有连接到电缆的站点把正在传送的帧字节存入它们的内部缓存器。帧的前六字节包含目的地址。站点在帧头识别出它的地址并把帧内容存入其内部缓存器，其他站点在这个阶段停止接收帧。目的节点处理收到的数据，沿着协议栈将它们上传，并通过电缆发送应答帧。以太网的帧包含目的地址和源地址；因此，接收端知道向哪里发送应答帧。

在节点1传输帧期间，节点2也尝试发送它的帧。但是，当它检测到介质正忙，因为有载波频率出现。这样，节点2必须等到节点1完成传输它的帧。

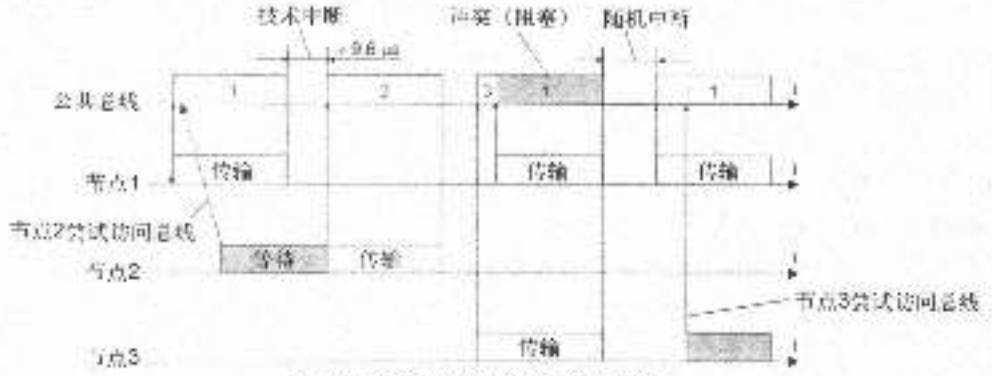


图12-7　CSMA/CD随机介质访问方法

完成帧传输后，所有的网络节点必须有技术中断，分组间距（interpacket gap，IPG），它持续9.6μs，这个中断是必须的以使网络适配器回到初始状态，防止一个站点独占介质。当IPG消失，网络节点有权传输它们的帧，因为介质是空闲的。图12-7的例子中，节点2等待节点1完成帧传输，并且经过9.6μs中断后开始传输它的帧。

12.3.3　冲突

载波侦听和帧间插入中断并不能确保消除两个或多个站点同时发现介质空闲并发送帧。这种情况就是冲突（collision），同时传送的帧内容在公共电缆中发生碰撞，结果，所有帧的信息被损坏，因为以太网采用的编码方法不能将每个站点的信号从公共信号中分离出来。

冲突在以太网操作中很普遍。图12-8的例子中，冲突源于节点3和节点1同时传输数据。冲突的发生并不一定是由几个节点同时传输造成；相反，这种情况是很少的。很有可能是一个节点开始传输帧，然后已经监视了介质的第二个节点并未发现载波（因为第一个节点发送的信号还没到达它）开始传输帧。因此，冲突是由网络节点的分布式定位引起的。

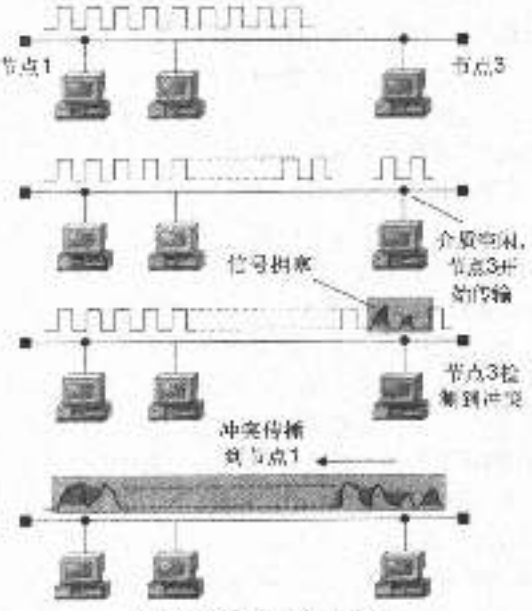


图12-8　冲突发生和传播

为正确处理冲突，所有站点同时检测电缆中的信号，如果发送信号和接收信号不一致，**冲突检测**（collisiondetection，CD）就发生。为提高所有网络节点快速检测到冲突的概率，站点检测到冲突后停止传送帧（在任意点，不一定是字节边界）并通过发送一个32位的序列到网络来增强冲突，这个序列就是**阻塞序列**（jam sequence）。

站点检测到冲突后必须停止传输并等待一个短暂的、随机的间隔，此后，它就能重新尝试获得介质并传送帧。随机中断的选择按下列算法：

$$\text{中断}=L\times(\text{时间槽}) \tag{12.1}$$

以太网中，时间槽在512位间隔（bt）里选取。位间隔对应于数据的两个序列位在电缆中出现的时间，对于10Mb/s以太网，位间隔是0.1μs，或者100ns。

L代表在0，2^N范围内等概率选取的一个整数，N是试图重传这个帧的重复次数：1，2，3，…，10。

第10次尝试后，中断的间隔就被固定并不在增加。这样，以太网技术随机中断的范围是0到52.4ms。

如果16次连续尝试传输帧引起了一个冲突，那么发送端就必须放弃并丢掉该帧。以上描述的算法就是**截断二进制指数退避**（truncated binary exponential backoff）。

当分析M/M/1模式的排队理论，高负载条件下（介质利用率增长并接近1时）的以太网工作情况对应第7章的图。尽管如此，以太网中访问介质的等待时间比在M/M/1模式中增长的更快：因为M/M/1模式太简单且没有考虑以太网的重要特性，如冲突。

基于共享介质的以太网管理者使用简单的经验法则，介质利用率不得超过30%。为支持延迟敏感业务，以太网（和其他基于共享介质的网络）只能使用一个QoS支持方法，也就是低负载操作模式（*underlaoded mode of operation*）。

12.3.4 路径延迟值和冲突检测

所有网络工作站对冲突的可靠检测是以太网正常运行的必备条件。如果传送工作站没有检测到冲突并判定数据帧被正确传送，那么这个帧就会丢失。因为在冲突条件下，帧信息会失真。由于校验和不匹配，接收工作站点将丢弃失真的帧。没有发送到接收端的数据可能通过一些面向连接的高层协议重新传输（如运输层或应用层协议或LLC协议，假设接收端操作在LLC2模式）。但是，通过高层协议的消息重传比通过以太网的消息重传耗费更多时间（有时达几秒），而后者操作只须微秒级的间隔。因此，以太网的网络节点不能可靠地检测冲突，那么就会明显减少网络的有效带宽。

为确保可靠的冲突检测，必须满足以下条件：

$$T_{\min} \geqslant \text{PDV} \tag{12.2}$$

这里，$T_{\min}$是传输一个最小长度的帧所需的时间，PDV代表路径延迟值，它表示冲突信号传播到最远网络节点的时间。最坏的情况下，信号必须在两个相隔最远距离的网络节点间通过两次。这种情况下，未失真的信号经过一个方向，由于冲突而失真的信号朝反方向传输。

如果条件满足（公式12.2），传送工作站必须完成帧传送之前，有时间能够检测由于它发送帧所导致的冲突。

这个条件的满足取决最小帧长度、协议的信息速率、网络电缆的长度和信号在电缆中的传播速率。这个速率对不同类型的电缆稍有不同。

以太网协议的所有参数的选择都是为了在网络节点的正常操作情况下，确保可靠的冲突检测。

以太网标准定义的最小数据字段长度为46字节，加上附属字段，产生一个64字节的最小帧长度；加上前同步码共72字节，或576位。

因此，可以估计出计算工作站之间的距离限制。在10Mb/s的以太网标准中，传送一个最小长度帧的时间需要575bt。所以，PDV必须小于57.5μs。这个时间内信号能传输的距离依赖于电缆类型：对于粗同轴电缆，大概是13 280m。考虑到这个时间内，信号必须经过通信链路两次，由此两节点间的距离的不能超过6 635m。考虑到其他情况，标准有十分严格的限制使这个距离更短。

一种限制与信号所允许的最大衰减有关。当信号经过两个独立且距离最远的站点时，为确保所需信号功率，并考虑到产生的信号衰减，粗同轴电缆连续段的最大长度是500m。很明显，在500m的电缆中，可以得到正确冲突检测的条件，对任意标准长度的帧都有很高的可靠性，包括72字节。500m电缆段的PDV仅为43.3bt。因此，可以使最小帧长度更小。然而，这种技术的研发者并没有减少它，因为他们心中有更加复杂的网络，有些包含了由多个中继器连接的多个段。

中继器增强了段与段之间传送信号的功率。结果，信号衰减减少，允许使用多个段增加网

络的长度。在用同轴电缆实现的以太网中，研发者限制网段的最大数量为5；这就相应限制了网络的总长度为2 500m。甚至在这样长的多段网络中，仍然满足冲突检测的条件并有所保留。例如，比较两个距离，一个是基于最大允许信号衰减得到的2 500m，另一个是依据信号传播时间计算得到的6 635m。实际上，因为在多网段网络中保留时间很短，中继器本身会给信号传播引入几十个比特位间隔的附加延时。自然地，为补偿电缆和中继器的参数偏差，也要考虑小的预留时间。

综合考虑这些因素，最小帧长度与两个网络工作站间的最大可能距离之间的关系应仔细选择。这种关系确保可靠的冲突检测。两个网络节点间的最大距离也称为**最大网络直径**。对所有类型的以太网，这个距离不能超过2 500m。

随着同样是基于CSMA/CD介质访问方法的新标准的帧传输速率特性的增长，例如快速以太网，网络站点间的最大距离按比例减少以增加传输速率：快速以太网中，它大约是210m；千兆以太网中，如果不是研发者决定增加最小分组长度，它将限制在25m以内。

表12-1提供了IEEE 802.3中详细说明的帧传输的主要参数值。这些参数不取决于物理介质，注意到以太网技术的物理介质的每个参数都将引入补充的、通常是更严格的限制。必须考虑这些限制，本章稍后将讨论它们。

表12-1　以太网MAC层参数

参　数	值
比特率	10Mb/s
时间槽	512bt
分组间距（IPG）	9.6μsec
尝试重传的最大次数	16
中断增加的最大数量	10
拥塞序列的长度	32位
最大帧长度（无前同步码）	1 518字节
最小帧长度（无前同步码）	64字节（512位）
前同步码长度	64位
冲突后随机中断的最小长度	0bt
冲突后随机中断的最大长度	524 000bt
工作站间的最大距离	2 500m
网络中工作站的最大数量	1 024

12.4　以太网帧格式

IEEE 802.3定义的以太网标准提供了MAC层帧格式。因为在IEEE 802.2中，MAC层帧必须包含LLC层帧，因此按照IEEE标准，以太网只能使用一种数据链路层帧，帧头是MAC子层和LLC子层帧头的组合。

然而，实际上，以太网使用四种帧格式。相同的帧类型可以有不同的名字，下列是一些最流行帧的名字。

- 以太网帧的第一个版本—**以太网 DIX/以太网 II**（Ethernet DIX/Ethernet II）出现在1980年，是三个公司共同努力的结果：美国数字设备公司、英特尔公司和富士施乐公司。这三个公司的联盟把其专用的以太网标准版本提交给IEEE 802.3委员会并将它确立为国际标准项目。
- 但是，802.3工作组通过的标准与DIX的建议在细节上有些不同；这些差异在于帧格式。因此，以太网帧的第二种形式出现：**802.3/LLC（802.3/802.2，或者Novell802.2）**

- 以太网帧的第三种形式——**原始802.3/Novell802.3**——它是Novell努力加快它的专用栈在以太网中运行的结果。
- 最后，帧格式的第四种形式：**以太网 SNAP**（Ethernet SNAP）（SNAP代表子网访问协议）。

这是IEEE 802.3委员会工作的结果，目的是确保遵从公共标准保持灵活性以满足未来添加字段和改变用途的需要。

帧格式的差异会导致网络硬件不兼容或软件只能在一种以太网帧格式下操作。然而，今天，实际上所有的网络适配器和它们的驱动器、网桥、交换机或路由器都能执行所有用到的以太网帧格式。必要的识别操作是自动执行的。

四种类型的以太网帧格式如图12-9所示。

802.3/LLC帧

6	6	2	1	1	1(2)	46–1497 (1496)	4
DA	SA	L	DSAP	SSAP	控制	数据	FCS
			LLC头部				

原始802.3/Novell 802.3帧

6	6	2	46–1500	4
DA	SA	L	数据	FCS

以太网DIX(II)帧

6	6	2	46–1500	4
DA	SA	T	数据	FCS

以太网SNAP帧

6	6	2	1	1	1	3	2	46–1492	4
DA	SA	L	DSAP	SSAP	控制	OUI	T	数据	FCS
			AA	AA	03	000000			
			LLC头部			SNAP头部			

图12-9 以太网帧格式

12.4.1 802.3/LLC

802.3/LLC帧的头部结合了IEEE 802.3和IEEE 802.2标准定义的帧头部字段。

802.3标准定义了8个头部字段（前同步码和帧起始定界符没有在图12-9中显示）：

- *前同步码*包含7个下列形式的同步字节：10101010。当使用曼彻斯特编码时，这种组合在物理层用频率为5MHz的周期性的波信号表示。
- *帧起始定界符*包含下列形式的单字节：10101011。这种位组合出现表明接下来的字节是帧头部的第一个字节。
- *目的地址*（DA）是2字节或6字节的长度。实际上，经常使用6字节的MAC地址。
- *源地址*（SA）是一个2或6字节字段，包含发送端的MAC地址。地址的第一位总置为零。
- *长度*（L）是2字节字段，决定帧的数据字段的长度。
- *数据*字段包含0到1 500个字节。但是，如果字段长度小于46字节，那么下一个字段是填充字段，用于补足上一帧使其达到46字节的最小可接收长度。
- *填充*字段包含足够多的填充字节以确保数据字段的最小长度：46字节。确保冲突检测机制的正确执行。如果数据字段的长度足够，帧中就没有填充字段。
- *帧校验序列*（FCS）由4字节组成，包含校验和。这个值按CRC-32机制计算得到。

802.3帧代表MAC子层的帧；因此，按照802.2标准，它的数据字段封装了去掉帧起始和结束标记的LLC子层的帧。LLC帧格式前面已介绍过。因为LLC帧有3字节或4字节（LLC2模式）的头

部长度（LLC1模式），数据字段的最大长度减少为1 497字节或1 496字节。

12.4.2　原始802.3/Novell 802.3帧

*原始*802.3（*Raw*802.3），也称为Novell 802.3，如图12-9所示，从中可以清楚看到这是一个按照802.3标准没有封装的MAC子层帧，很长时间内，Novell在它的NetWare操作系统中不使用LLC帧的辅助字段。没必要标明封装在数据字段中的信息类型，因为它包含在IPX分组中。很长一段时间，IPX是Novell NetWare中仅有的网络层协议。

当需要标明出现的高层协议时，Novell采用把LLC帧封装在MAC层帧中的形式（如，它使用标准802.3/LLC帧）。Novell公司指出在它的操作系统中使用像802.2帧格式的帧，尽管它们实际上是802.3和802.2头部的组合。

12.4.3　以太网DIX/以太网II帧

*以太网DIX*又称为*以太网*Ⅱ，其结构与原始802.3帧结构相一致（见图12-9）。然而，原始802.3帧的2字节L字段用作*以太网DIX*帧的协议类型字段。这个字段称为*类型*（*Type*，*T*）或*以太类型*（*EtherType*），作用与LLC帧的*DSAP*和*SSAP*字段目的是一样的，也就是说，具体指明高层协议的类型，它的分组被封装到这个帧的数据字段。

对比SAP字段的协议代码，它是1字节长度，*T*字段提供2字节的协议代码。因此，相同的协议在SAP和*T*字段被编码为不同的数值。例如，IP中*以太类型*（*EtherType*）字段的代码是2048_{10}（0x0800），而在SAP字段中，同样的协议编码为6。以太类型字段的协议代码值出现在SAP值之前，因为以太网DIX专有版本出现在采用802.3标准之前。所以，当符合802.3的设备普及的时代到来时，这些数值对大多数硬件和软件产品已经是事实上的标准。因为以太网DIX和原始802.3的结构一致，*长度/类型*字段在文件中通常表示为*L/T*字段。这个字段的数字决定了它的用法：如果这个值小于1 500，则它是*L*字段；否则，是*T*字段。

12.4.4　以太网SNAP帧

为消除封装在以太网帧数据字段中的协议类型信息的编码矛盾，IEEE 802.2工作组努力将以太网的帧进一步标准化。结果，新的以太网帧出现：以太网SNAP（见图12-9），一种802.3/LLC帧的扩展。这种扩展是通过引入附加的、包含两个字段SNAP头部实现的：*OUI和类型*（*Type*，*T*）。*T*字段包含2个字节并且是仿照以太网Ⅱ帧中*T*字段的格式和用途。这意味着此字段使用和协议代码相同的值。*OUI*字段定义了在*T*字段中控制协议代码的组织标识符。引入的SNAP头部与以太网Ⅱ帧的协议代码能够兼容并且创造了一种通用的协议编码方法。802技术的协议代码由IEEE控制，它的*OUI*值是000000。如果某些新技术可能需要其他的协议代码，对于分配这些代码的组织指定其他*OUI*值足够了；旧代码值将仍然有效（具有适当的OIU）。

因为SNAP是封装在LLC协议中的一种协议，所以DSAP和SSAP字段包含了分配给SNAP协议的代码0xAA。LLC头部的*控制*（*control*）字段置为0x03，对应使用未编号的帧。

SNAP头部是对LLC头部的补充，因此它在以太帧和其他802技术的帧中都是允许的。例如，IP总是在将它的分组封装进所有LAN协议帧的过程中使用LLC/SNAP头部结构，这些LAN协议是：FDDI、令牌环网、100VG-AnyLAN、以太网、快速以太网和千兆以太网。还有，当使用以太网、快速以太网和千兆以太网网络传送IP分组时，IP使用以太网DIX帧。

12.4.5　使用各种类型的以太网帧

因为有四种类型的以太网帧，所以网络层协议必须解决选择具体帧类型的问题。必须做出选择——总是只使用一种帧类型、使用所有4种类型或者优先选择某个具体的帧类型。

IP使用两种类型的帧：原始以太网Ⅱ或以太网SNAP，后者有更复杂的结构。IP优先选用以太网Ⅱ帧。

现代以太网适配器能够依据帧字段值自动识别以太网帧的类型。例如，以太网Ⅱ帧通过长度/类型字段值很容易从其他类型字段中区分出来。如果，这个值超过1 500，它就是类型字段，由于协议代码的值总被选择超过1 500。类型字段出现表明这是一个以太网Ⅱ帧，它唯一地在给定位置使用这种字段。

IPX能工作在所有的以太网帧类型。它能使用以上描述的方法识别以太网的帧，如果所讨论的帧有其他类型，*L/T*字段的值小于或等于1 500，则要做更进一步的检测。对帧类型的进一步识别可通过判断LLC字段是否出现来实现。仅当L字段紧跟在IPX分组的起始点（即2字节）后时，LLC字段才可以省略。这个字段是全1的，值为0xFFFF，或是两个设置为255的字节序列。首先，试图像DSAP和SSAP字段一样解释这两个字节。但是，*DSAP*和*SSAP*字段不可能同时含有这个值，因此2个字节置为255表明这是原始802.3帧。

其他的情况，可以依据*DSAP*和*SSAP*的值做深入的分析。如果它们都置为0xAA，则它们是以太网SNAP帧；否则，就是802.3/LLC帧。

12.5 以太网的最好性能

网络性能取决于帧经过通信链路的传输速率和通信设备处理它们的速率。当处理帧时，通信设备在与通信链路相连的端口之间传送它们。帧经过通行链路的传输速率依赖所使用的物理和数据层协议。例如，有可能是以太网10Mb/s、以太网100Mb/s、令牌环网和FDDI。

协议经过通信链路传送位的速率称为**标称协议速率**（nominal protocol rate）。

通信设备处理帧的速率依赖设备处理器的性能、内部构造和其他参数。显然，通信设备的性能必须与链路的传输速率相一致。如果低于链路速率，帧在序列中被延迟并且当缓存满时被丢弃。另一方面，使用的通信设备的性能参数高于通信链路几百倍是没有任何意义的。

为了评估配备有以太网端口的通信设备的要求性能，必须评估以太网段（*Ethernet egment*）的性能。这种评估不应以b/s来进行（正如我们所知，这个值等于10Mb/s），而应该以帧数每秒来进行。这是因为网桥、路由器或交换机几乎同时处理每一个帧，不管它的长度：所需的时间包括查找前向表，形成新的帧（用路由器）等等。另一方面，当帧的长度最小时，每个时间单位到达的帧的数量自然就最大。因此，通信设备最难的操作模式是**处理具有最小长度的帧流**（processing a flow of frames having a minimum length）。

说明 当详细说明网络的性能时，术语，帧和分组是同义的，因此，度量单元如帧每秒（f/s）和分组每秒（p/s）也是相同的。

使用表12-1提供的参数，用每秒传送的具有最小长度的帧（分组）的数量来计算以太网段的最大性能。

在计算可以通过以太网段的最小长度的帧的最大数量之前，必须注意最小长度帧的大小应该加上前同步码，所以最小长度帧是72字节，数据段的最小长度是46字节。因此，帧的最小长度是576位（图12-10），传输它需要57.5μs。加上IPG（9.6μs）结果是：67.1μs。这样，以太网段的最大可能吞吐量（*the maximum possible thoughput of the Ethernet segment*）是14 880fps。自然地，网段内有多个节点时，这个值会减少，因为节点等待直到被允许访问介质所需要的时间和冲突的出现。

以太网中最大长度的帧有一个1 500字节的数据字段。包括辅助信息，共有1 518字节；处理最大长度的帧时以太网段的最大吞吐量是813fps，当处理长帧时，网桥、交换机和路由器的负载明显

减少。

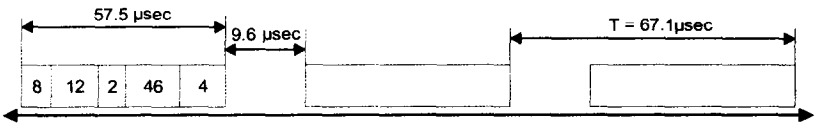


图12-10　计算以太网协议的吞吐量

现在，使用不同长度的帧计算一个网段的最大有效带宽。(用位每秒为单位)

有效协议带宽是帧的数据字段所携带的用户数据的最大传输速率。

这个带宽总是小于以太网协议的标称位速率，原因如下：

- 帧中控制信息的存在
- 分组间距
- 访问介质的等待时间

对于最小长度的帧，有效带宽是：

$$B_e=14\ 880\times 46\times 8=5.48\text{Mb/s} \tag{12.3}$$

这个值小于10Mb/s，应当考虑到最小长度的帧主要用于传输确认信号。这个速率与文件数据的传输速率关系不大。

对于最大长度的帧，有效带宽是：

$$B_e=813\times 1\ 500\times 8=9.76\text{Mb/s}$$

当使用中等长度的帧时，数据字段512字节，协议带宽是9.29Mb/s。

在后两种情况下，协议带宽被证明相当接近于10Mb/s的最大带宽，尽管执行这种评估时，认为没有其他站点干扰两个通信站点间的通信（如，没有冲突且不考虑要等待访问介质)。

因此，当没有冲突时，网络利用率取决于帧的数据字段长度并且传输最大长度的帧时其值最大是0.976。

12.6　以太网物理介质规范

历史上，第一个以太网络使用直径为0.5英寸的同轴电缆。此后，以太网标准定义了其他物理层规范，允许使用各种数据传输介质。对于10Mb/s以太网的物理介质规范，CSMA/CD和所有的时间参数是一样的。

以太网物理规范包含以下数据传输介质：

- 10Base-5——直径0.5英寸的同轴电缆，又称为粗同轴电缆。阻抗是50欧姆。最大段长（没有中继器）是500m。
- 10Base-2——直径0.25英寸的同轴电缆，又称为细同轴电缆。阻抗是50欧姆。最大段长（没有中继器）是185m。
- 10Base-T——非屏蔽双绞线电缆（UTP)。由一个中央集中器组成的星型拓扑结构。集中器和端节点间的距离不得超过100m。
- 10Base-F——光纤电缆，其拓扑结构类似于10Base-T。这个规范有几个版本：光纤中继器间链路，或FOIRL（距离可达1 000m)；10Base-FL（距离可达2 000m）和10Base-FB（距离可达2 000m)。

规范名称里面的10代表按照10Mb/s标准数据传输的比特速率。Base是指使用10MHz作为基本频率（基带）的编码方法（和使用多种载波频率的方法相比，又称为宽带)。标准最后的参数是指

电缆类型。

12.6.1 10Base-5

10Base-5标准通常和Xerox建立的实验以太网相对应，被认为是经典的以太网标准。

图12-11显示了基于粗同轴电缆网络的不同组件和包含由中继器连接的三个网段。

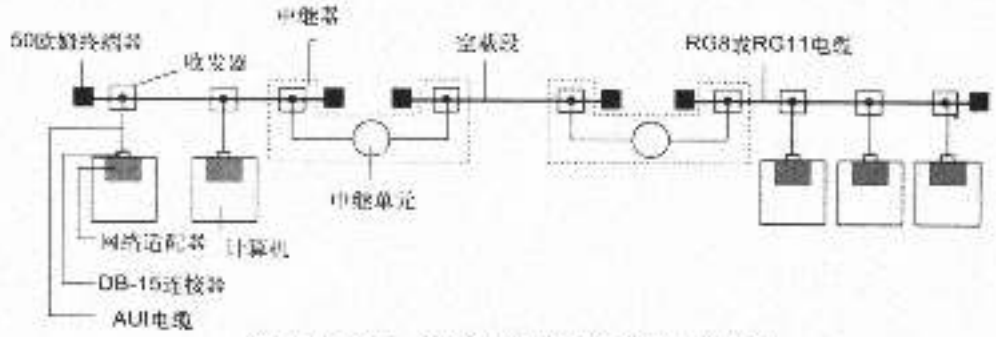


图12-11 10Base-5网络物理层组件组成的一个网段

所有站点的电缆用作单信道。一个电缆段的最大长度是500m（没有中继器），必须在两个端节点之间用50欧姆的终端器（terminator）连接。终端器吸收信号通过电缆后的能量使它们不反射到链路上。如果没有终端器，电缆中会出现固定的波形信号，因此，有些站点会收到强信号，发送到其他站点的信号很微弱以至于接收不到它们。

工作站必须使用收发器（transerver）连接到电缆，收发器是网络适配器的一部分，执行传送和接收的功能（发送端+接收端=收发器）。收发器直接和电缆相连并由嵌入在计算机中的网络适配器提供能量，收发器可以通过穿过电缆的T型头或者通过非接触的方法连接到电缆，确保直接和物理接触（插入式分接头）。

收发器通过附加单元接口（attachment unit interface，AUI）电缆连接到网络适配器，接口电缆可以达到50m。AUI包含四对双绞线（网络适配器必须有一个AUI连接器）。当一种类型的电缆转到另一种类型时，收发器和网络适配器的保留部分之间的标准接口是非常有用的。为此目的，可以替代收发器，网络适配器的保留部分不需要替代，因为它执行MAC层协议。这样，只需确保新的收发器（例如，用于双绞线的收发器）支持标准AUI。

一个单独的网段不能超过100个收发器，收发器连接点的距离不能小于2.5m。电缆每2.5m需做标记表明连接收发器的点。当计算机按照这些标记点连接的，电缆中驻波对网络适配器的影响就会降到最小。

简化的收发器结构如图12-12所示。收发器和接收端使用一种特殊的电路（如，转换电路）连接到电缆的同一点。这种电路采用的结构可以同时接收或发送到达和来自电缆的信号。

如果适配器发生故障，可能发生任意信号序列不断通过电缆的情况，因为对所有工作站，电缆是共享介质，一个有故障的适配器将会妨碍网络操作。为避免这种情况，收发器的输出端安装了一个特殊的装置能够检验帧传输时间。如果超过帧传输的最大可能时间（考虑一些预留时间），这些装置能够将电缆与收发器的输出端断开。帧的最大传输时间（包括前同步码）是1 221μs，帧传输的限制是4 000μs（4ms），这种收发器的功能有时称为超长控制（*jabber control*）。

冲突检测器（collision detector）通过增加信号直流成分的电平发现同轴电缆中的冲突。如果直流成分超过特定的阈值（大约1.5V），这意味着至少两个收发器同时传送信号到电缆。

去耦元件（Decoupling elements）确保收发器对来自网络适配器保留部分的电流去耦合，这样能保护网络适配器和计算机在损坏的电缆上免受高电压骤降的影响。

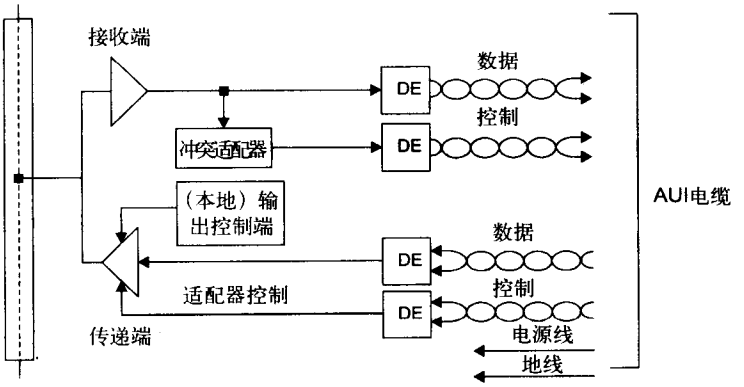


图12-12　收发器结构设计

10Base-5决定可以使用一种特殊设备又称为中继器。中继器将几个电缆段连接成一个单独的网络，这样就增加了网络的总长度。中继器接收来自电缆段的信号并同步地按位将它们转发到其他网段，改善信号形状，增加信号的功率和同步脉冲。一个中继器包含两个（或更多）连接到电缆段的收发器，和自身具备时钟振荡器的转发单元。为了更好的同步，中继器延迟传输帧前同步码的第一位，增加段到段的帧传输延时并减小IGP。

标准允许网络中有不超过4个中继器，因此，电缆不应超过5段。给定电缆段的最大长度（500m），就给出了10Base-5网络的最大长度是2 500m。这与以太网最大网络直径的一般限制相对应。

五个段中只有三个能加负载（也就是端节点连接到它们）。负载段必须与空载段分开，因此，最大网络结构是连接着空载段的两个负载段与中间的负载段相连。这章的前面，图12-11中，有个以太网的例子，包含了两个中继器连接的三个网段。边缘段是负载，中间段是空载。

以太网10Base-5网络使用中继器遵守的规则又称为**5-4-3规则**：五个段、四个中继器、三个负载段。

中继器的数量限制是由于中继器引入的附加信号传播延时所决定的。中继器的使用增加了PVD，可靠的冲突检测不能超过传输最大长度的帧所要求的时间（也就是，1帧包含72字节，或576位）。

每个中继器通过自己的收发器与段连接，因此不超过99个节点，而不是100个，可以与负载段相连。这样，一个10Base-5网络中端节点的最大数量是99 × 3=279个节点。

12.6.2　10Base-2

10Base-2标准使用同轴电缆作为传输介质，同轴电缆的中央铜线的直径是0.89mm，外部直径大约是5mm（细以太网）。电缆阻抗是50欧姆（ohm）。没有中继器的最大段长度是185m；阻抗为50欧姆的终端器必须连接段的端节点。细同轴电缆比粗同轴电缆便宜，因此10Base-2网络有时又称为廉价网。尽管如此，低价格有它的不利之处，因为细同轴电缆易受噪声干扰、较低的机械强度、信号带宽更窄。

连接电缆的工作站使用BNC **T型连接器**（T-connector），它是T型连接的，一个搭线头连接网络适配器；另外两个搭线头连接电缆。可以连接单个段的最大工作站数量是30。工作站间的最小距离是1m。细同轴电缆的特点是以1m的步长连接端节点。

10Base-2也规定按照5-4-3规则使用中继器。

这样，网络的最大长度为5 × 185=925m。很明显，这比2 500m的一般限制更严格。

说明 建立正确运行的以太网网络时要遵守许多限制，有些限制和网络参数有关，例如网络的最大长度或计算机的最大数量，必须同时满足多个条件，正确的以太网网络必须满足所有要求。实际上，仅满足最严格的限制条件就足够了。例如，一般限制规定以太网不能超过1 024个节点，10Base-2标准限制连接单个段的最大站点数量是30，负载段的限制数量为3个，因此，10Base-2网络最大节点数量不超过29×3=87。

10Base-2与10Base-5类似。但是在10Base-2中，收发器集成在网络适配器里。由于非常柔软的细同轴电缆可以直接连到BNC T型连接器，它装在计算机网络接口卡的背面。电缆连着网络适配器，使得移动计算机变得容易。

图12-13可以看到，基于10Base-2并且包含两个单独电缆段的典型网络。

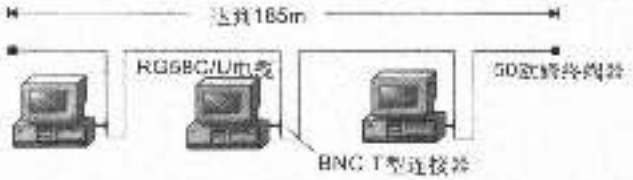


图12-13 10Base-2标准网络

这种标准的实际执行导致电缆网络的最简单的解决方案，仅需要一个网络适配器、T型连接器和50欧姆的终端器把计算机连接到网络。但是，这种电缆连接类型易受到攻击而使电缆发生故障；细以太网电缆比粗同轴电缆对噪声更加敏感。单倍道有许多机械连接（每个T型连接器有三个机械连接，其中两个对整个网络是十分重要的）。用户能够访问连接器并能打乱单倍道的完整性。这种方案按照美学和环境学要求都不算完美，因为，电缆中两个非常明显的部分使用T型连接器连接每个站点。桌子下面会形成很乱的区域，因为必须提供空地以防止工作场所轻微移动。

10Base-5和10Base-2标准的共同缺点是缺乏关于单倍道状态的联机信息。由于网络不能操作时，电缆故障可以被立刻检测到，但是需要一个特殊的称为电缆测试器的设备发现电缆部分的故障。

12.6.3 10Base-T

10Base-T使用两对非屏蔽双绞线（UTP）作为传输介质。电话公司使用基于3类UTP的多对电缆连接建筑物里的电话机有很长时间了。这种电缆有个形象的名字——语音级别——表明它主要用于语音传输。

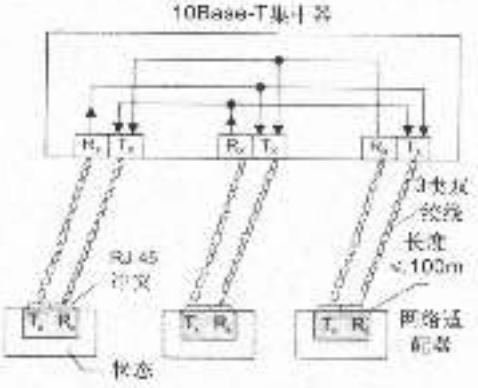


图12-14 10Base-T标准网，T_i作为传送端，R_i作为接收端

运用这种流行电缆建立LAN的想法证明是非常有成果的。因为大多数建筑物配有必须的电缆系统，它被保留以便用到一种方法能够把网络适配器和其他通信设备连接到双绞线，与以太网使用的同轴电缆相比，这种做法使网络适配器和网络操作系统的通信软件变化最小。这个问题已经成功解决；切换到双绞线仅需要替换网络适配器中的收发器或路由器端口，访问的方法和数据链路层的所有协议与同轴电缆以太网是相同的。

按照“点到点”的拓扑结构，两对双绞线连接端节点到一个称为多端口中继器的特殊设备。一对双绞线用于将数据从站点传输到中继器（网络适配器的T_i输出端），另一对双绞线将数据从中继器传输到站点（网络适配器的R_i输入端）。图12-14显示了三

个端口的中继器的例子。中继器接收来自一个端节点的信号，同时将信号传送到除信号接收端口之外的所有其他端口。

多端口的中继器通常在技术术语上又称为**集中器**（concentrator），或**集线器**（hub）。集中器对连接到其端口的双绞线部分执行信号中继器的功能，这就形成了公共数据传输介质：一条逻辑单信道（或逻辑公共总线）。一旦当同时有信号传输到集中器的多个RI输入端时，集中器就检测到段中的冲突。在这种情况下，集中器发送扰乱序列到它所有的TI输出端。此标准定义了数据传输速率为10Mb/s、直接连接节点（站点和集中器）的双绞线部分的最大长度不超过100m，假设使用3类或更好的双绞线。这个距离由双绞线的带宽决定，这个带宽允许使用曼彻斯特编码在100m的距离以10Mb/s的速率传输数据。

10Base-1集中器可以使用相同的端口互连，这些端口用于连接端节点。这样做时，必须确保发送端和接收端的一个端口连接，相应地，接收端和发送端的另一个端口连接。

10Base-T标准定义网络中任意两个站点间集中器的最大数量为4个，这就是**4集线器规则**（4 hubs rule）。

4集线器规则代替了用于同轴电缆的5-4-3规则。它能确保执行CSMA/CD访问程序的工作站点之间的同步并确保可靠的冲突检测。

当建立的10Base-T网络有大量的站点时，集中器使用分层方法互联，这样就形成一个树状结构（图12-15）。

说明　10Base-T不允许集中器环形连接，因为它将导致错误的网络操作。这个要求意味着10Base-T不允许重要的集中器之间有平行链路，平行链路中，如果端口、集中器或电缆失效就会产生链路限制。10Base-T网络的链路限制可能是平行链路的一条连到阻塞（停止）的站点。

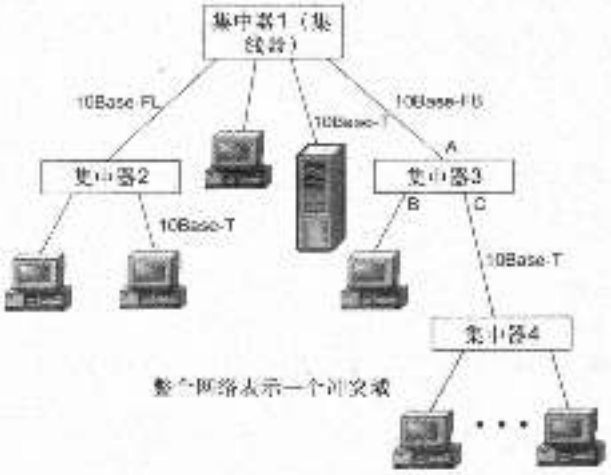


图12-15　以太网集中器的分层连接

10Base-T网络中工作站的最大数量不超过1 024个的一般限制。对于这种类型的物理层，这个限制是可以达到的：完全能创建两层的集中器层次，在其底层放置足够多的集中器使总端口数为1 024个（图12-16），端节点必须连接低层集中器的端口。这种情况下，4集线器规则能满足，因为任意两个端节点间有3个集线器。

由于任意两个工作站点间不能多于4个中继器，很明显，10Base-T网络的最大网络直径是5×100=500m。注意这个限制比以太网中2 500m的一般限制要严格得多。

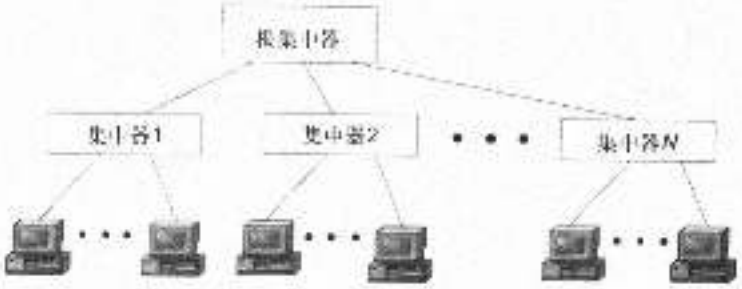


图12-16 最大数量的工作站系统

建立在10Base-T基础上的网络比同轴电缆以太网有更多的优势，因为公共物理电缆被分成了多个单独的电缆部分连接到中央通信设备。尽管逻辑上这些部分继续形成共享介质，但电缆的物理分离意味着它们能被控制并且如果网络适配器损坏、短路或故障的，能单独地分开电缆。这能明显简化大型以太网的维护程序，因为集中器通常自动执行这些操作并通报网络管理员可能出现的问题。

10Base-T定义了测试连接端点收发器和中继器端口的两对双绞线段的物理可操作性的程序。这个程序称为链路完整性测试（*link integrity test*）。它是基于在每对双绞线的接收端和发送端之间以16ms的时间间隔传送特殊信号（曼彻斯特码的*J*和*K*）。曼彻斯特编码的信息信号在一个时钟的中间总是改变电压，*J*和*K*通过在时钟的中间维持电压来破坏这种规则。两个电压值中的一个对应*J*码，另一个值对应*K*码。由于*J*和*K*在帧传输期间无效，测试序列对介质访问算法的操作没有影响。

端节点间引入有源设备使其能控制节点的操作并能从网络隔离不正确操作的节点。这是10Base-T技术比同轴网络的优势之处，后者难于维护，基于有集中器，以太网具备基本的容错能力。

12.6.4 光纤以太网

10Base-F网络使用光纤电缆作共享介质。光纤标准推荐相对低廉的多模光纤作为主要的电缆类型。这种光纤经过1km的电缆长度的带宽是500MHz-800MHz。较贵的单模光纤带宽有几千兆是可以接受的，尽管必须使用特殊类型的收发器。

基于光纤的以太网包含和10Base-T相同的元件，即网络适配器、多端口中继器和适配器连接到中继器端口的电缆部分，如使用双绞线，两根光纤用于连接适配器到中继器：一根光纤连接适配器的T_x输出端到中继器的R_x输入端，另一根连接适配器的R_x输入端到中继器T_x的输出端。

光纤中继器间链路（Fiber Optic Inter-Repeater Link，FOIRL）是IEEE 802.3委员会为以太网中使用光纤制定的第一个标准。它保证中继器间的光纤链路的长度是1km，假设总的网络长度不超过2 500m，任意两个网络节点间的最大中继器的数量是4个。这样，就能实现最大直径2 500m。尽管全部4个中继器同及中继器和端节点间的电缆的最大长度都不允许达到，否则，产生的网络直径可达5 000m。

10Base-FL标准是对FOIRL做了小的改进，增加了发送端的功率，因此，端节点和集中器间的最大距离上升到2 000m。节点间的中继器的最大数量仍为4个，网络最大长度是2 500m。

10Base-FB标准仅考虑连接中继器，端节点不能使用这种标准连接集中器端口。网络节点间可以连接5个10Base-FB中继器，单个段的最大长度是2 000m，网络最大长度是2 740m。

按照10Base-FB连接的中继器在没有帧发送时，它们不断地转发不同于数据帧信号的特殊信号序列。这样做以便维持同步。因此，当段与段间传送数据时会引入较小的延时。这是只允许5个中继器的最大原因。曼彻斯特码用做按以下顺序的特殊信号：J J K K J J……这个序列产生2.5MHz的脉冲，使一个集中器的接收端和另一个集中器的发送端同步。因此，10Base-FB标准又称为**同步以太网**（synchronous Ethernet）。

正如10Base-T一样，光纤以太网标准允许集中器连接成树状层次结构。集中器端口间不允许有环路。

在这章的开始，10Base-F作为专业术语用于所有三种10Mb/s以太网光纤标准。它不是标准术语，但网络专家有时使用它作为通用昵称。

12.6.5 冲突域

冲突域是以太网网络的一部分，在它上面的节点可以检测出一个冲突，而检测过程与冲突发生所在的网络位置无关。

建立在中继器基础上的以太网总会形成一个冲突域。网桥、交换机和路由器将以太网网络分成多个冲突域。

如图12-15所示的网络是个单冲突域。例如，如果集中器4中有一帧冲突，那么按照10Base-T集中器的操作逻辑，冲突信号将传播到所有集中器的所有端口。

另一方面，如果集中器3被网桥替代，那么它的连接集中器4的端口C，将接收到冲突信号但不会把它传送到所有其他端口，因为这超出了它的能力。网桥使用端口C对冲突进行简单处理，端口C连接到冲突发生的共享介质。如果冲突发生时网桥试图通过端口C传送一帧到集中器4，那么已经记录了冲突信号，端口C将停止传输帧并在一个随机时间间隔后试图重传帧。如果端口C接收帧时发生冲突，它将简单地丢弃收到的碎片并等待，直到发送帧的节点通过集中器4重新传输帧。当成功把帧收到缓存器后，网桥把它传送到另一个端口，如端口A，如转发表中所规定。

由端口C处理的与冲突相关的所有事件对所有连接到网桥其他端口的所有网段都是未知的。

12.6.6 10Mb/s以太网标准的公共特性

表12-2和表12-3总结了10Mb/s以太网标准的主要限制和特性。

表12-2　所有以太网标准的公共限制

特　性	值
额定带宽	10Mb/s
网络中工作站点的最大数量	1 024
网络节点间的最大距离	2 500m (对于10Base-FB是2 750m)
网络中同轴段的最大数量	5

表12-3　以太网物理层规格参数

参数	10Base-5	10Base-2	10Base-T	10Base-F
电缆	粗同轴RG-8 或RG-11电缆	RG-58细 同轴电缆	3、4、5类 UTP	多模光纤 电缆
最大段长度（m）	500	185	100	2 000
网络节点间的最大距离（使用中继器）（m）	2 500	925	500	2 500 (2 740对于10Base-FB)
单段工作站点间的最大数量	100	30	1 024	1 024
两个工作站点间的最大中继器数量	4	4	4	4(5对于10Base-FB)

12.7 案例学习

20世纪90年代初，大型工程技术工厂Transmash使用10Mb/s的共享介质以太网将其所有小型机

和个人电脑互联在一起（图12-17）。计算机主要用于执行自动任务，它们之间的数据转发非常少。网络传送少量由数字和字母组成的数据，因此一个公共的共享介质足以满足生产需求。基于10Base-FB和10Base-FL标准的光纤网络用于连接网络的中央段和远处车间的段。网络满足以太网结构的所有要求：电缆部分不超过最大允许长度，任意两节点间不多于4个集线器，网络节点间的最大距离不超过1 800m（图12-17中的计算机A和计算机C）。

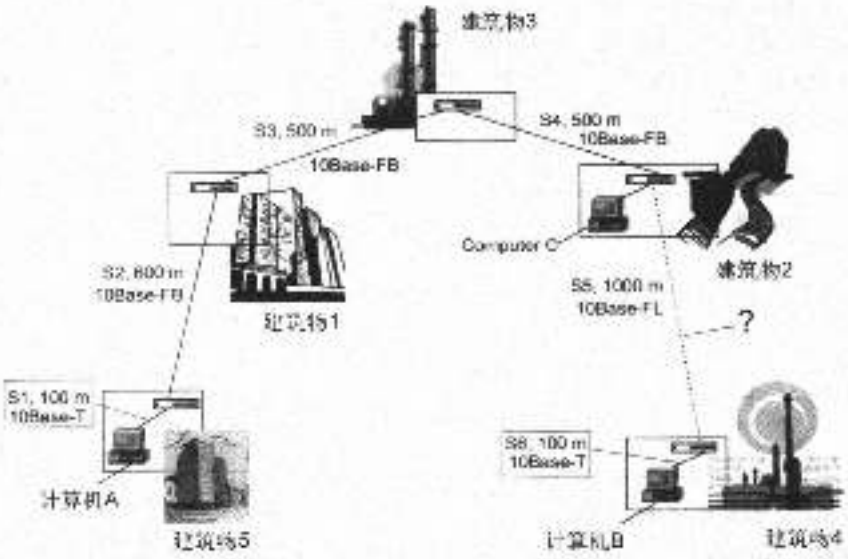


图12-17 Transmash的多段以太网网络

一段时间后，需要连接另一个建筑4到网络。这个建筑坐落在使用光纤以太网标准（10Base-FB或10Base-FL）连接的网络范围以内。尽管如此，这种连接会产生错误的结构，因为建筑物1和建筑物4的计算机之间会有5个集线器。此外，网络直径将达到2 800m，这又违背以太网的另一个限制。但是，此时Transmash的网络架构师不想从根本上改变网络结构，他可以安装网桥和路由器连接新的段。他知道IEEE 802.3标准的第13节，题为“多段10Mb/s基带网络的系统考量”，提供了评估网络结构正确性的程序。这种技术可以定量地决定特定的网络结构是否将正常操作。计算显示有时它可能违反4集线器规则和最大网络直径的限制，但保护了正确的结构。这种限制的选择是为确保留有安全余度。例如，对于任何节点的可靠冲突检测，PDV的最大值不能超过575bt。这意味着10Base-5网络（4个集线器和网络直径是2 500m）最大结构有38bt的预留。同时，802.3标准的第13节提供的模式指出甚至4bt的预留就能使网络正确操作。

因此，网络架构师计算Transmash工程技术工厂添加一新段的可能结构。已经证实，建筑物4连接到网络，网络将有6.6bt的预留。重新检查计算后，可以安装光纤电缆，连接建筑物4到工厂网络，且新的网络结构能够运转。实际证明计算是正确的，整个网络运行正常。这样的结构保持了几年，直到新的应用需求增长导致必须把公共共享介质划分成交换段。

为检验Transmash网络架构师实现的计算结果，有必要让你熟悉802.3标准的第13节提供的程序细节。

这个程序描述了当下列条件满足时，以太网将正常操作：

- 两个最远工作站点的PDV不超过575bt。中继器和设备所在信号传播中引入附加延时。IEEE 802.3标准的第13节的表提供了延时的阈值数据。
- 减少IPG和顺序列消息通过所有中继器后的路径可变值（PVV）不超过49bt。每个中继器通

过特定值减少IPG，这在标准的第13节的表中提供。

802.3标准的表提供了可能的信号传播延时和IPG的减少量，因为它们的具体值取决于中继器的制造商。Transmash的网络架构师根据网络设备制造商提供的精确数据进行计算。表12-4和表12-5提供了这些数据。

考虑如何使用表12-4提供的数据计算PDV。

802.3标准的研发者试图尽可能简化计算，因此表12-5提供的数据包含了信号传播的几个阶段。例如，中继器引入的延时包括收发器输入端引入的延时，中继器单元引入的延时和收发器输出端引入的延时。然而，在表中所有这些延时用单个值表示，称为*段基*（*segment base*）。

为避免再次加上电缆引入的延时，对每种电缆表中都提供了双倍延时值。

表12-4　计算PDV的数据

段类型	左段基（bt）	中间段基（bt）	右段基（bt）	介质延时每米（bt）	最大段长度（m）
10Base-5	11.8	46.5	169.5	0.0866	500
10Base-2	11.8	46.5	169.5	0.1026	185
10Base-T	15.3	42.0	165.0	0.113	100
10Base-FB	—	24.0	—	0.1	2 000
10Base-FL	12.3	33.5	156.5	0.1	2 000
FOIRL	7.8	29.0	152.0	0.1	1 000
AUI(>2m)	0	0	0	0.1026	2 048

表中使用定义如*左段*（*left segment*）、*右段*（*right segment*）和*中间段*（*intermediate*）。让我们为Transmash工程技术工厂阐明这些术语（图12-17）。目的是在最坏的情况下计算PDV，假定选择节点*A*和*B*，它们被5个中继器隔开且它们之间的距离是2 800m。

802.3术语中，左段是从发送端端节点的输出端开始的信号路径。术语与左段所在的地理（或它们在图中）位置无关。只是对开始计算段的习惯称呼。选择S1段，与节点A线连，就是左网段。

此后，信号通过中间段S2-S5并到达连接段S6的接收端（节点B），表中假设最坏的情况是这点发生冲突。冲突发生的末尾段称为右段。

每段都会引入固定延时，称为**基（base）**，它仅依赖于段的类型和它在信号路径中的位置（左、中间或右）。冲突发生的右段的基明显超过左段和中间段的基。

此外，每段都引入信号传播延时，这个延时依赖段的长度并且通过计算信号在1m电缆的传播时间（位间隔）乘以电缆长度（单位米）得到。

计算还包括每个电缆部分引入的估计延时（表中提供长度以便计算电缆每米的信号延时乘以段长度）。然后，所有这些延时被加到左段、中间段和右段的基。

因为左段和右段有不同的基延时值，那么如果网络边缘处段的类型不同，就必须计算两次。首先，对左段的一种段类型进行评估并对左段的另一种段类型重新计算。最大的PDV值作为最终的结果。这样，边缘段是相同的类型，即10Base-T；因此，重复计算没有必要。

现在计算PDV：

- 左段S1：
  15.3（基）+（100×0.113）=26.6
- 中间段S2：
  24+（600×0.1）=84.0
- 中间段S3：
  24+（500×0.1）=74.0

- 中间段S4：
  24+（500×0.1）=74.0
- 中间段S5：
  33.5+（1 000×0.1）=133.5
- 右段S6：
  165+（100×0.113）=176.3

这些值相加得出PDV值为568.4。

因为PDV值小于6.6bt允许的最大值575，所以网络结构是正确的，尽管它的总长度超过2 500m并且多于4个中继器。

但是，检验PDV不足以做出正确的结论：必须评估PVV。

表12-5提供了PVV计算的初始值。

表12-5 通过中继器减少IPG

段类型	传送段（bt）	中间段（bt）
10Base-5或10Base-2	16	11
10Base-FB	—	2
10Base-FL	10.5	8
10Base-T	10.5	8

按照这些信息，计算PVV：

- 左段1 10Base-T：10.5bt
- 中间段2 10Base-FL：8
- 中间段3 10Base-FB：2
- 中间段4 10Base-FB：2
- 中间段5 10Base-FB：2

这些值的总和得出PVV为24.5，比49bt的阈值小。

因此，可以推断出符合以太网标准的所有参数，包括段长度和中继器数量。

小结

- 共享LAN是最简单和廉价的LAN实现类型。LAN主要缺点是低可靠性，因为随着节点数的增加，分配给每个节点的带宽相应地减少。
- IEEE 802委员会开发的标准包括LAN底层的推荐设计：物理和数据链路层。LAN的特殊性反映在把数据链路层分为两子层：LLC和MAC。
- MAC子层负责访问介质和用于发送帧。IEEE 802标准使用的各种方法分为两类：随机的和确定性的。随机访问方法确保在低介质负载情况下介质访问的延时最小。但是，随着介质利用率接近100%，随机访问方法的使用导致大量延时。确定性方法能够在高网络负载条件下运行。
- 802.1工作组标准对所有技术都共用。它们定义了LAN类型、它们的特性、网络互联的程序以及网桥和路由器的操作逻辑。
- LLC协议确保为高层协议的运输服务所要求的质量。它能使用数据报传输或者使用建立连接的方式发送帧和存储帧。
- 当今，以太网是最普通和广泛使用的LAN技术。广义的以太网是一个技术族，包括专用以太网DIX标准和开放标准，如IEEE 802.3的10Mb/s以太网、快速以太网、千兆以太网和10G

以太网。除了10G以太网以外，所有类型的以太网使用相同的访问方法——CSMA/CD——它在许多方面决定了技术性质。

- 冲突是以太网的重要事件。它发生在两个站点同时通过共享介质传送数据帧时。冲突的发生是以太网的自然属性，是使用随机访问方法产生的必然结果。可靠的冲突检测依赖于网络参数的正确选择。尤其靠观察最小帧长度和最大网络直径之间的关系。
- 传送最小长度的帧时，以太网10Mb/s段可达到的最大吞吐量用帧每秒表示为：14 880f/s。
- 以太网带宽的最大可能值是9.75Mb/s，对应使用1 518字节最大长度的帧以513 f/s传送。
- 以太网支持4种帧类型，有共同的节点地址形式。网络适配器按形式化法则自动识别帧类型。
- 按照物理介质类型，IEEE 802.3定义不同的形式：10Base-2、10Base-T、FOIRL、10Basee-FB。对每种具体形式，定义下列特性：电缆类型、连续电缆段的最大长度和使用中继器增加网络直径的规则，即同轴网络的5-4-3规则以及双绞线和光纤的4集线器规则。

复习题

1. 使用以太网的例子解释网络可延拓性和网络扩展性之间的不同？
2. 比较访问介质的确定性方法和随机方法？
3. 为什么广域网技术的数据链路层不分为MAC和LLC子层？
4. LLC执行什么功能？
5. 什么是冲突？
6. 以太网标准中的前同步码和帧起始定界符的功能是什么？
7. 执行拥塞控制的网络工具是哪个？
8. 为什么以太网引入分组间距？
9. 10Base-5标准的下列特性值是多少？
   - 额定带宽（b/s）
   - 有效带宽（v/s）
   - 吞吐量（f/s）
   - 分组内传输率（b/s）
   - 位间间隔（sec）
10. 10Base-5标准为何把帧的最小长度设置为64字节？
11. 为什么10Base-T和10Base-FL/FB使得实际上不再使用同轴以太网标准？
12. 解释以太网帧每个字段的意义？
13. 以太网帧格式有4种形式。从下面提供的清单中选出这些标准的名字，考虑一些标准有多个名字：
    a. Novell 802.2
    b. 以太网 Ⅱ
    c. 802.3/802.2
    d. Novell 802.3
    e. 原始802.3
    f. 以太网 DIX
    g. 802.3/LLC
    h. 以太网 SNAP
14. 当不同格式的以太网帧被传送到网络会发生什么情况？
15. 分组的大小值怎样影响网络？对太长的帧会有什么问题？为什么短帧效率低？
16. 利用率如何影响以太网网络性能？

17. 共享介质以太网网络的数据传输率如何影响最大网络直径？
18. 以太网中，选择物理段的最大长度要考虑哪些因素？
19. 从FOIRL标准转换到10Base-FL期间，什么使段长度的最大值增加？
20. 引起4集线器规则的这个限制的原因是什么？
21. 集中器为什么不支持全双工以太网模式？

练习题

1. 图12-18中的网络部分表示冲突域吗？

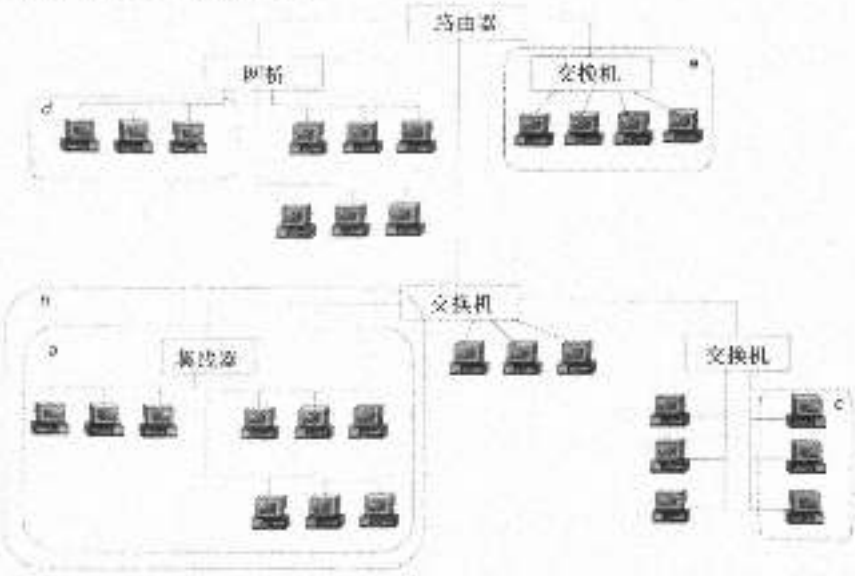


图12-18 可能的冲突域

2. 站点在它的帧被网络适配器丢弃之前要等多久？
3. 如果建立在集中器基础上的网络出现封闭环（环路）会发生什么？
如图12-19所示的例子。

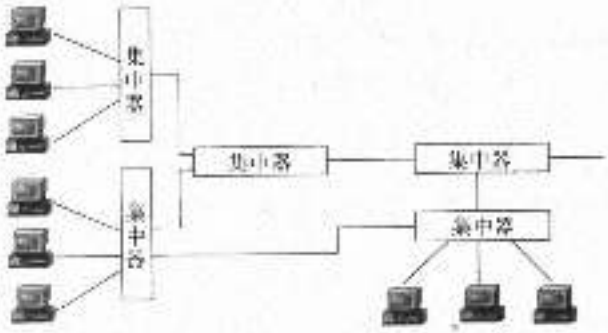


图12-19 基于集中器的以太网中的环路

a. 网络正常运行。
b. 帧将不能到达目的节点。
c. 试图发送任何帧时都会发生冲突。
d. 帧将循环。

4. 如果丢失和损坏帧的比率从0%升到3%，那么当传送240 000字节的文件时，评估以太网网络的性能下降。如图12-20说明的网络操作。

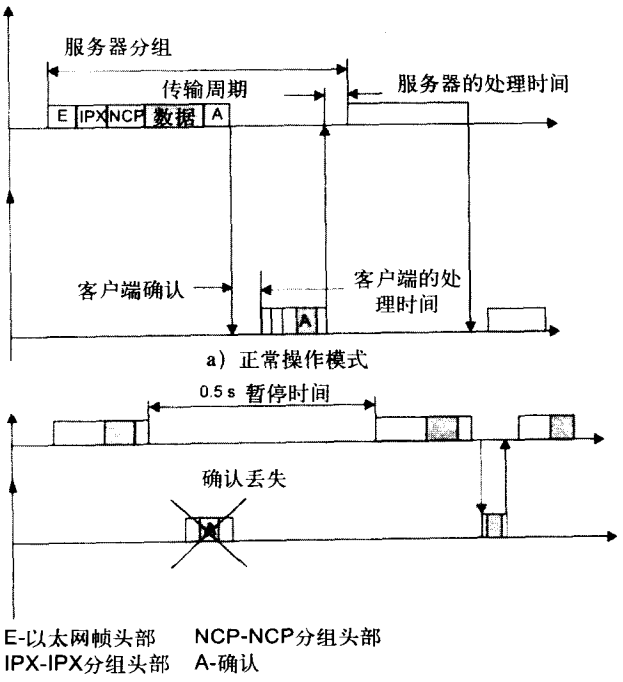


图12-20　文件传输过程中以太网网络操作

使用下列协议传送文件：以太网、IPX（网络层）和NCP（文件服务的应用层）。协议头部的大小如下：

- 以太网——26字节（包括前同步码和FCS字段）
- IPX——30字节
- NCP——20字节

文件以1 000字节段传输。只有按空闲源方法操作的NCP存储丢失或损坏的帧。等待有效确认的暂停时间固定为500ms。（这不是NCP操作的唯一模式：它也能使用滑动窗口算法操作，但是这里没有使用此模式）。确认信号的大小是10字节。客户端处理一个分组的时间是650μs；服务器端是50μs。

提示　这个问题包含两部分。首先必须确认在理想网络操作条件下文件传输的实际速率，此时丢失和损坏以太网帧的比例为零。问题的第二部分确认帧开始丢失和损坏时的文件传输速率。

文件传输总共需要240个分组。携带被传输文件的1 000字节的以太网帧的大小是1 000+20+30+26=1 076字节或8 608位。

携带确认信号的以太网帧的大小是86字节（加上前同步码）或688位。

在这些条件下，理想网络中传输文件下一部分的时间周期是860.8+68.8+650+50=1 629.6ms。

传送240 000字节所需时间是240 × 1 629.6=0.391s，信息速率为240 000/0.391=613 810b/s。现在，剩下的是找出帧开始丢失或损坏时的信息速率。

第13章 高速以太网

13.1 引言

十五年来，经典10Mb/s以太网能够满足大多数用户的要求，但是在20世纪90年代早期，它的带宽不足的缺点开始显现。10Mb/s的交换速率相对于内部计算机总线速率是非常小的，后者当时已经超过1 000Mb/s的阈值（PCI总线确保数据传输率为133MB/s）。这将导致使用PCI总线的服务器和工作站在网络上的运行变慢。

对新的以太网技术的需求更加迫切。这种新技术在性价比方面应当与以前的技术效率相同，并且确保100Mb/s的性能。在研究和调查的过程中，专家分成两组，最后导致1995年出现两种新技术：快速以太网和100VG-AnyLAN。但是，最后只有快速以太网存活下来，它保留了更多的经典以太网性能，包括CSMA/CD。

快速以太网的成功更增加了人们对高速以太网的兴趣。下一代的变体——千兆以太网三年后被标准化。千兆以太网的成功同样是因为继承了10Mb/s以太网的高性能；它也保留了基于CSMA/CD共享介质的操作。

但是，以太网最新的版本——10G以太网与前者有较大的不同。特别是，它只执行全双工的模式，这意味着不再支持共享介质。

因此，本章只考虑快速以太网和千兆以太网。10G以太网将在第15章介绍，包括运行在全双工模式从而能够构建交换LAN的其他技术。

13.2 快速以太网

13.2.1 历史概述

1992年，一个网络设备制造集团，包括以太网领导者如SynOptics公司和3Com公司，创建了非商业的*快速以太网联盟*（*Fast Ethernet Alliance*）。这个联盟的目标是为新技术研发标准，这种技术必须有相当大的性能提高，同时应尽可能多地保留以太网的性质。

那时，IEEE 802委员会已经成立了研究组调查新的高速技术的潜能。从1992年底到1993年底，IEEE工作组研究了不同厂商提供的多种100Mbit的解决方案。和快速以太网联盟的建议一样，工作组也在考虑Hewlett-Packard公司和AT&T公司建议的技术。

保留CSMA/CD的问题是讨论的中心。快速以太网联盟建议保留这种方法，这样能保证10Mb/s和100Mb/s技术的兼容性和一致性。HP/AT&联合体支持它的网络设备制造商数量少于支持快速以太网联盟的数量，提出了一种新的不同的访问方法称为**需求优先**（demand priority）。这种方法极大地改变了网络节点的行为模式并且不能与以太网和IEEE 802.3融合。一个新的IEEE工作组（IEEE 802.12）对需求优先方法进行标准化。

1995年秋天，上述两种技术都成为IEEE的标准。IEEE委员会采纳快速以太网作为802.3u标准。IEEE802.3u并不是独立的标准，而是对现存的802.3标准的补充，802.3标准的内容在第21到30章。802.12委员会采纳100VG-AnyLAN，它使用新的需求优先访问方法并且支持两种格式的帧：以太网和令牌环。

13.2.2　快速以太网的物理层

快速以太网和经典以太网的所有不同集中在物理层（图13-1）。在快速以太网中，MAC和LLC层保持不变并且在802.3和802.2标准的同一章节中描述。因此，对于快速以太网，这里只考虑其物理层的几个版本。

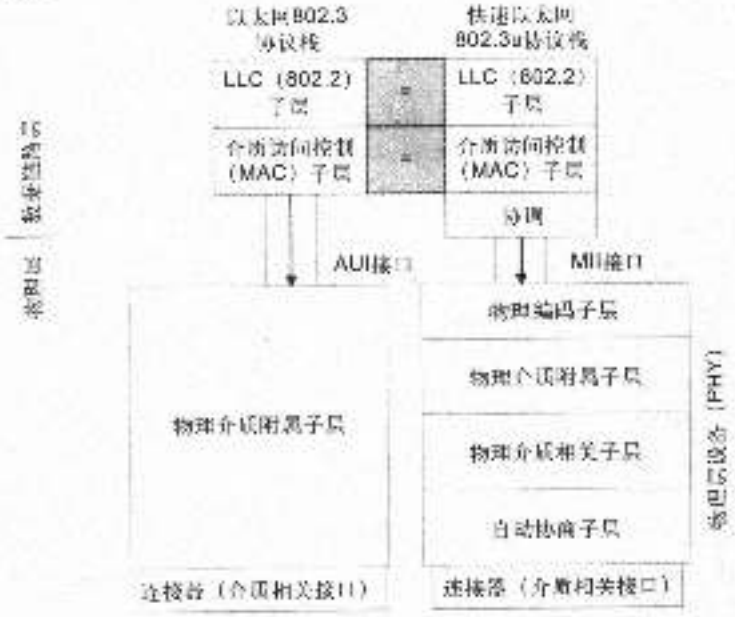


图13-1　快速以太网和经典以太网的不同

快速以太网的结构比较复杂，是因为使用了三种不同的电缆系统：

- 光纤多模电缆，两根光纤。
- 5类双绞线，两对。
- 3类双绞线，四对。

世界最早以太网网络使用的同轴电缆并不包括在快速以太网所允许的传输介质清单中。这是大多数新技术的共同趋势，因为5类双绞线可以使数据传输率与同轴电缆相同。同时，网络成本更低并且维护更方便。当距离非常长时，光纤提供的带宽远大于同轴电缆，网络成本只是略有提高，尤其是考虑到大型同轴电缆系统中寻找和排除故障的高花费，这就不算什么了。

由于不使用同轴电缆，基于共享介质的快速以太网通常为以集中器基础的层级树状结构，这与10Base-T/10Base-F网络相似。快速以太网结构的最大不同是网络直径减少到大约200m，和经典以太网相比，数据传输率提高10倍，传输最小长度帧的时间减少10倍。

然而，这并不表示建立基于快速以太网的大型网络会有许多障碍。20世纪90年代中期的标志事件是广泛使用廉价的高速技术和交换LAN（*switched* LAN）的快速发展。使用交换技术，快速以太网可以执行全双工模式，对网络总长度没有限制。使用全双工模式的唯一限制是连接相邻设备的物理段长度（如适配器—交换机或者交换机—交换机）。

本节将集中讨论经典的快速以太网的半双工形式，完全对应802.3标准定义的访问方法。第15章将考虑执行全双工模式的快速以太网的具体性质。

与经典以太网的物理实现（有六个）相比，快速以太网中各种物理层变体之间的不同更重要。这是因为在这些实现之间导体数量和编码方法都不同。另外，快速以太网的多种物理实现是同时建立的而不是像经典以太网那样是逐步演变而来的。因此，可以具体定义物理层的子层，而且不

会随各种物理媒体的具体版本和子层的不同而发生改变。

官方802.3标准定义了快速以太网物理层的三种规范：

- 100Base-TX用于基于5类非屏蔽双绞线（UTP）的两对电缆或者1类屏蔽双绞线（STP）。
- 100Base-T4用于基于3、4、5类UTP的四对电缆。
- 100Base-FX用于使用两根光纤的多模光纤电缆。

以下对所有三种规范都有效：

- 快速以太网帧格式与10Mb/s以太网使用的帧格式相同。
- 分组间距（IPG）是0.96μs，位间隔（bt）是10ns。访问算法的所有定时参数（时间槽时间、传输最小长度帧的时间等等）都用位间隔描述并且保持不变，因此MAC层部分的标准没有发生变化。
- 介质中传输有适当冗余码的空闲符号表明介质的可用性（10Mb/s以太网标准没有信号指名介质空闲）。

物理层包含三个元件：

- **介质独立接口**（Media independent interface，MII）。
- **协调子层**（reconciliation sublayer）。确保MAC层能够操作AUI，为了与AUI一起运行，协调子层确保MAC层与使用MII的物理层进行通信。
- **物理层设备**（Physical layer device，PHY）依次包含以下多个子层（图13-1）。
  - **物理编码子层**（Physical coding sublayer，PCS）将来自MAC层的字节转换成4B/5B 或8B/6T符号（两种编码都用在快速以太网中）。
  - **物理介质附属**（Physical medium attachment，PMA）子层和**物理介质相关**（Physical medium dependent，PMD）子层确保电或光信号的形成如NRZI或MLT-3。
  - **自动协商子层**（Autonegotiation sublayer）允许两个相互作用的端口自动选择最有效的操作模式，例如，半双工和全双工。这层是可选的。

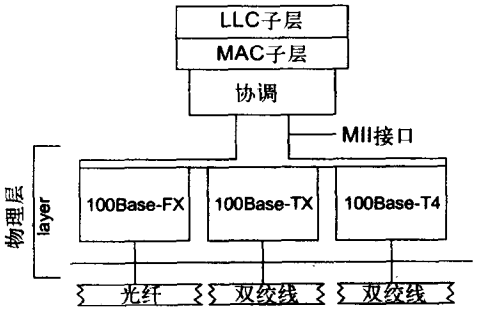


图13-2 快速以太网物理层的结构

MII使用介质独立方法在MAC和PHY子层间交换数据。为此目的，这个接口与经典以太网的AUI相似，只是AUI处在物理层信令（PLS）子层（不同的电缆使用相同的物理编码方法，曼彻斯特码）和PMA子层中间。另一方面，MII处在MAC层的协调子层和PCS之间，支持如前面所提到的两种编码方法。PCS、PMA和PMD形成了PHY子层，快速以太网的物理子层三种版本：FX、TX和T4（图13-2）。

13.2.3 100Base-FX/TX/T4 规范

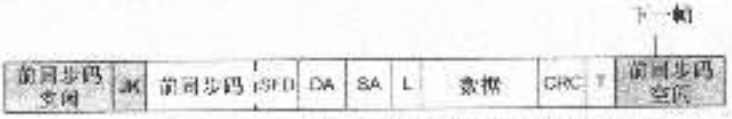
100Base-FX、100Base-TX和100Base-T4规范有许多共同点。因此，我们将在广义范围讨论这些规范的共同性质，例如，100Base-FX/TX或者100Base-TX/T4。

100Base-FX规范（多模光纤，两根光纤）定义以半双工和全双工模式运行的快速以太网协议。

当通过电缆传输数据时，10Mb/s以太网使用曼彻斯特码表示数据，与之相比，快速以太网定义了另一种编码方法：4B/5B。4B/5B编码的细节已在第9章讨论过了。研发快速以太网之前，FDDI网络就显示了这种方法的有效性并且不加改变地将其引入100Base-FX/TX之中。这种方法中，每4位MAC子层数据（也称符号）用5位表示。当用电和光脉冲的形式表示5位信元时，见余位可以使用潜在的码元。

因为4B/5B和8B/6T冗余位能够丢弃错误信号。非法编码的存在可以提高100Base-FX/TX网络的稳定性。因此，在快速以太网中，表示介质能够使用可以通过重复传输空闲符号（*Idle symbol*）（11111）。即当对用户数据编码时有一个非法编码。这种方法使接收端总是与发送端同步。

为了从空闲符号中分离出以太网帧，使用起始定界符。这包含两个符号：4B/5B编码的J（11000）和K（10001）。帧传输完成后，将T符号插入到第一个空闲符号前（图13-3）。



JK——有意义符号流的起始点定界符
T——有意义符号流的结束点定界符

图13-3　100Base-FX/TX的连续数据流

当MAC编码的4位块转化成物理层的5位块后，它们必须用光或电信号表示并传送到与网络节点连接的电缆。100Base-FX和100Base-TX使用不同的线性编码方法，分别是NRZI和MLT-3。

*100Base-TX*规范使用5类UTP或1类STP电缆（两对）作为传输介质。

它与100Base-FX规范最大的不同是使用MLT-3方法传送信号——4B/5B编码的5位块——使用双绞线和自动协商功能选择端口的操作模式。

自动协商（autonegotiation）方法允许两个物理连接的并且支持不同物理层协议的设备，根据位速率和双绞线对数量支持不同的物理层标准，按照选择最有效的操作模式。通常，当能够用10Mb/s和100Mb/s的网络适配器连接交换机或路由器时，自动协商过程将发生。

基于双绞线的100Base-TX/T4设备支持五种操作模式：

- 10Base-T
- 10Base-T 全双工
- 10Base-TX
- 10Base-T4
- 100Base-TX 全双工

在协商过程中，10Base-T的优先级最低，全双工100Base-TX模式的优先级最高。

协商过程发生在设备启动后，它也能在设备的控制单元启动时发生。

自动协商过程的设备发给它的对方一个特殊的脉冲序列，称为**快速链路脉冲串**（fast link pulse burst，FLP），包含一个8位数字，它从当前节点支持的最高优先级模式开始，对被推荐的作用模式编码。

如果对方节点支持协商功能并且支持建议的模式，它用另一个FLP串应答，确认建议的模式。这时协商过程完成。如果对方节点仅支持较低的操作模式，则在应答中指出这种模式。这种模式就被选作操作模式。这样，就选择了它们共同支持的最高优先级模式。

快速以太网的物理层规范出现后不久就产生了100Base-T4规范（3类UTP，四对）。10Base-TX/FX的早期技术开发者的主要目标是创建一个物理层规范，使其尽可能类似于10Base-T和10Base-F，它们在两种数据传输链路之间操作：两对双绞线或两根光纤。为在两对双绞线间执行操作，有必要更换为更高质量的5类电缆。

同时，与之竞争的100VG-AnyLAN技术开发者最初把注意力集中在使用3类双绞线。这种方案的主要优势并不是它的价格低而是这种电缆已经安装在大多数的建筑物中。因此，100Base-TX和100Base-FX发布后，快速以太网的开发者也使用3类双绞线作为其物理层版本。

取代4B/5B编码，这种方法使用有更窄信号频谱的8B/6T编码。在33Mb/s,使用3类双绞线，这种方法的频带固定为16MHz。（当使用4B/5B编码，信号的频谱不适用这个频带）。MAC子层信息

的每8位块用六个三元符号编码（即，数字有三种状态）。每个三元数字的传输持续40ns。六个三元数字独立而顺序地被传送到三对传输双绞线之一。

第四对总被用作侦听载波以检测冲突。前三对双绞线的每对传输率是33.3Mb/s，因此100Base-T4协议的总速率是100Mb/s。同时，因为采用的编码方法，每对电缆的信号变化率是25M波特，所以可以使用3类双绞线。

图13-4显示了100Base-T4 网络适配器的MDI（介质独立接口）端口和集中器MDI-X端口（MDI-X port）的连接。（后缀X表示这个端口中，连接接收端和发送端的插脚可以交换的，以便将一个设备发送的信号路由到另一个设备以便使其接收信号，反之亦然，这就提供了一个简单的方法连接电缆对而不用将它们交叉。）对1-2用于将数据由MDI端口传送到MDI-X端口，对3-6用于MDI端口接收来自MDI-X端口的数据，对4-5和7-8是双向的并用于传输和接收。

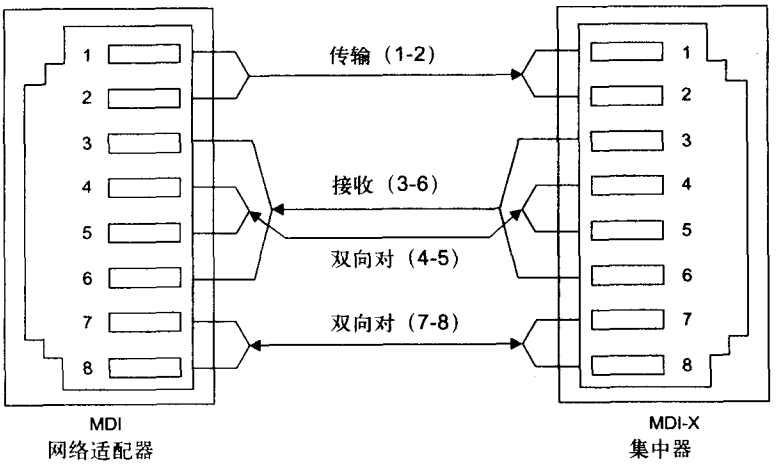


图13-4 100Base-T4中的节点连接

13.2.4 使用中继器构建快速以太网段的规则

像所有的以太网形式一样，快速以太网使用集中器和中继器创建网络链路。

正确构建快速以太网段的规则包括：

- 连接数据终端设备（DTE）之间的最大段长度的限制。
- 连接DTE和中继器端口之间的最大段长度的限制。
- 最大网络直径的限制。
- 中继器的最大数量和连接中继器的最大段长度的限制。

1. DTE-DTE段的最大长度的限制

网络中数据帧的任何源端，包括网络适配器、网桥或路由器端口、网络控制模块或其他类似设备，都能成为DTE。DTE设备的出众特性在于它能为共享段产生新的帧。（尽管网桥或路由器传输的帧是由网络适配器使用它们的输出端口产生的，但这些帧对连接到特殊输出端口的网络段来说是新的。）然而，中继器的端口不是DTE设备，因为它能将出现在段中的帧一位一位地复制。

典型的以太网配置中，多个DTE设备连接到中继器端口，从而形成网络的星形拓扑。DTE-DTE连接没有出现在共享段中（除非奇特的配置，两台计算机的适配器通过电缆直接连接）。另一方面，网桥或路由器这样连接是非常普通的。这种情况下，或者网络适配器直接连接一个设备的端口，或者这些设备直接相互连接。

按照IEEE 802.3μ，表13-1提供了最大的DTE-DTE段长度。

表13-1　最大的DTE-DTE段长度

标准	电缆类型	最大段长度
100Base-TX	5类UTP	100m
100Base-FX	多模光纤	412m（半双工）到
	62.5/125μm	2Km（全双工）
100Base-T4	3、4或5类UTP	100m

2. 基于中继器的快速以太网的限制

快速以太网中继器分为两类：

- I类中继器（class I repeater）支持所有类型的逻辑数据编码：4B/5B和8B/6T。这意味着I类中继器允许以100Mb/s的速率传送逻辑代码。因此，I类中继器具备所有三种物理层类型的端口：100Base-TX、100Base-FX和100Base-T4。
- II类中继器（Class II repeater）支持4B/5B或者8B/6T。II类中继器没有100Base-T4端口或100Base-TX和100Base-FX端口，因为这些物理层规范使用4B/5B代码。

一个冲突域只能有一个I类中继器，它引入了一个重要的信号传播延时，因为必须将信号由本地代码转换为其他代码。这个延时是70bt。

II类中继器引入了更小的信号传播延时：对TX/FX端口是33.5bt，对T4端口是46bt。因此，在单个冲突域中，II类中继器的最大数量是2个。

快速以太网中继器的限制并不表示建立大规模网络有一系列困难，因为可以使用交换机和路由器将网络划分成多个冲突域，每个域都基于一或两个中继器，这样总的网络长度没有限制。

表13-2列出了使用I类中继器建立网络的规则。

表13-2　使用I类中继器的快速以太网的参数

电缆类型	最大网络直径（m）	最大段长度（m）
仅用双绞线（TX）	200	100
仅用光纤（FX）	272	136
基于双绞线的多个段和基于光纤的一个段	260	100(TX)160(FX)
基于双绞线的多个段和基于光纤的多个段	272	100(TX)136(FX)

图13-5显示了典型网络配置的限制。

这样，将快速以太网的距离线器规则简化为一或二集线器规则（*one or two hubs rule*），集线器的数量取决于集线器的类型。

当确定网络配置正确时，不使用一或二集线器规则，可以计算路径延时值（PDV），正如第12章中10Mb/s的例子所显示的。

就像使用10Mb/s以太网一样，802.3标准提供了快速以太网中计算PDV的参考数据。

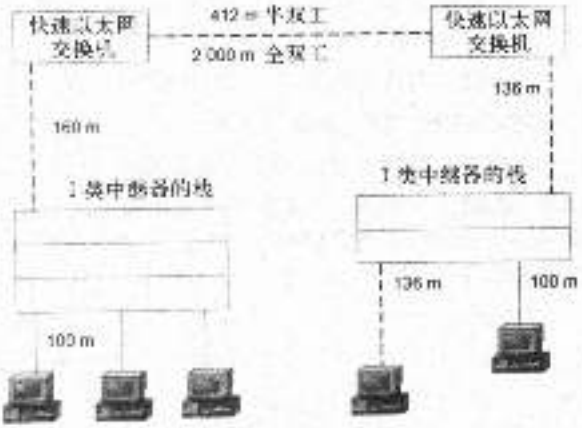


图13-5　使用I类中继器的快速以太网的例子

13.2.5　100VG-AnyLAN的特殊性质

尽管100VG-AnyLAN实现了许多先进的技术方案，但它并没有获得很多的支持并

且被淘汰。它没有找到它的应用领域，因为它与传统的更简单的快速以太网相比，它被证明是过于复杂。这点是毫无疑问的，因为千兆以太网支持的应用需要高的传输速率，确保数据传输率在1000Mb/s并且保留了以太网和快速以太网的历史链路。

与快速以太网相比，100VG-AnyLAN与经典以太网有许多方面的不同。

访问共享介质是基于一种原理上完全不同的方法——**需求优先**。这种访问方法是让集中器执行仲裁器的功能，解决访问共享介质的问题。100VG-AnyLAN网络包含一个**中央集中器**（central concentrator）也称为**根集中器**（root concentrator），将端节点和其他集中器连接起来（图13-6）。

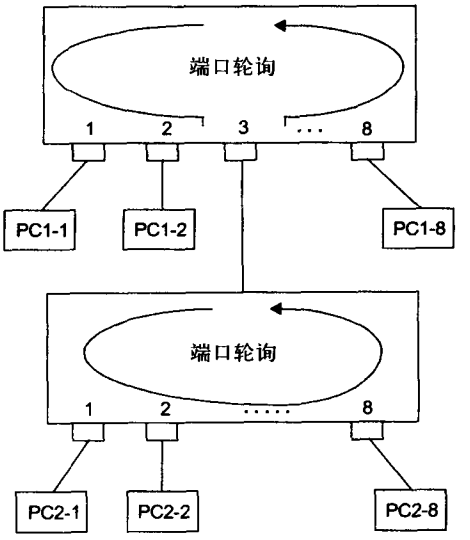


图13-6 100VG-AnyLAN网络

100VG-AnyLAN网络允许三种层次层叠。必须配备集中器和100VG-AnyLAN适配器以便操作以太网或令牌帧。不允许两种类型的帧同时循环。

集中器对所有的端口进行回合制轮询。需要发送分组的站点发出一个特别的低频信号给集中器，要求允许其传送帧并指定优先级。100VG-AnyLAN网络使用两种优先级：低和高。低优先级对应于一般的数据传输（文件服务、打印服务等等），高优先级对应于延时敏感数据（如多媒体）。请求优先级有静态和动态两种，这意味着如果低优先级站点长时间没有访问网络，那么自动给低优先级站点分配高优先级。

如果介质可用，那么集中器允许传输分组。当集中器分析收到分组的目的地址后，自动将分组发送到目的节点。如果网络忙，集中器就按照请求到达的先后，并考虑它们的优先级，把收到的请求放入队列中。如果有其他集中器连接到该端口，那么轮询被推迟直到低优先级的集中器完成轮询。连接不同层次集中器的站点都平等地访问共享介质，因为只有当所有集中器完成轮询所有端口后，才决定提供访问介质。

集中器如何知道连接目的站点的哪个端口被连接上？在所有其他技术中，帧被发送到所有其他网络站点，目的站点识别出帧的地址，将帧复制到它的缓冲器。为解决这个问题，当站点通过电缆物理地连接到网络时，集中器就识别出站点的MAC地址。与其他技术不同，通过物理连接测试电缆的连通性（10Base-T的链路完整性测试）并决定端口的运行速率（快速以太网的自动协商），100VG-AnyLAN集中器确定建立物理连接的站点的MAC地址并且将它存放在MAC地址表中（这个表类似于网桥或路由器表）。100VG-AnyLAN集中器和网桥或路由器的不同是集中器没有内部缓存器存储帧：因此，只能从网络中接收一帧，将它发送到目的端口，直到这个帧被目的站点完全接收后它才接收新的帧。这意味着保留了共享介质的效应。这种改进仅有利于增强网络的安全性，因为帧不发送到外部端口，使它们很难被检测到。

100VG-AnyLAN支持多种物理层规范。第一个版本是使用四对3，4和5类UTP。不久后其他形式出现，如两对5类UTP，两对1类STP，或两根多膜光纤。

13.3 千兆以太网

13.3.1 历史概述

快速以太网产品出现在市场不久，网络集成者和管理者在建立企业级网络时发现了一些问题。在许多情况下，通过100Mb/s信道互连的服务器会使得FDDI和快速以太网的主干网过载，这些主

千网的运行速率也是100Mb/s对下一代速率层次的需求愈发明显。1995年，只有ATM交换机能提供高的速率，但很少用于LAN，因为其成本高且与经典以太网技术有明显不同。

因此，IEEE的下一步工作非常合乎逻辑。1996年夏天，802.3z工作组建立：它目标是开发一种尽可能类似于以太网的协议，但能提供1 000Mb/s的位速率。就像快速以太网一样，这个消息得到了以太网支持者的热情接受。

这种热情的主要原因是网络主干可以平稳地转变到千兆以太网，类似从拥塞的低水平以太网段转变到快速以太网。另外，已经积累了以千兆速率传输数据的经验。在MAN和WAN中，这是通过SDH完成。在LAN中，光纤信道可以实现同样的目标。后者主要用于将高速外围设备连接到功能强大的计算机。确保通过光纤电缆的数据传输率接近1千兆；这可以使用冗余8B/10B代码实现。

8B/10B代码作为千兆以太网物理层的首个版本被用于光纤信道。

802.3z标准在1998年6月通过。使用5类双绞线运行千兆以太网的目标交给了802.3ab任务组。由于使用这种类型的电缆传输千兆数据速率的复杂性，最初支持的速率大约为100Mb/s。802.3ab任务组成功地实现了目标，使用5类双绞线的千兆以太网不久被采纳。

13.3.2　问题

千兆以太网开发者的主要目的是尽可能多地保留经典以太网的思想，同时达到1 000Mb/s的位速率。

尽管希望引入的技术创新从出现的新技术中反映技术进步的共同趋势是合理的，但这些期望并没有在千兆以太网中得到满足。尤其是，像它的低速处理器，千兆以太网不能支持协议层的下列性质：

- 质量服务（QoS）
- 冗余链路
- 网络节点和设备可用性测试（后种情况，正如10Base-T以太网、10Base-F和快速以太网，没有端口到端口的链路）

所有这三种功能对于当代网络尤其是不久后的网络都是有前途和有用的。为何千兆以太网的开发者放弃它们？

答案很简单。在当今的LAN中，交换机可以确保这些有用的性质，支持以太网系列协议的全双工版本。因此，快速以太网开发者决定基本的协议必须简单的确保快速数据传输，比较复杂的功能并不总是需要（如QoS支持）的，且必须由控制交换机的高层协议承担。

因此，千兆以太网和它前面的以太网和快速以太网有什么共同性质？

- 保留了所有以太网的帧格式
- 支持CSMA/CD协议的半双工版本仍然存在。保留基于共享介质的廉价方案，允许千兆以太网应用于有高速服务器和工作站的小型网络。
- 支持用于以太网和快速以太网的主要电缆类型，包括光纤、5类双绞线和STP。

尽管千兆以太网的开发者决定不创建先进的新性质，甚至在确保经典以太网的基本功能时，它们也面对许多复杂问题：

- 运行在共享介质上确保可接受的网络直径。因为CSMA/CD限制电缆长度、保留帧大小和CSMA/CD的所有参数，将共享介质形式的千兆以太网最大段长度减小到25m。因为许多应用领域需要网络直径至少为200m，所以必须找到解决方案并且不对快速以太网做较大改变。
- 使用光纤实现1 000Mb/s的位速率。光纤信道作为千兆以太网光纤版本的物理层基础，确保数据传输率为800Mb/s。
- 假设支持双绞线电缆。首先，这个问题看起来好像无法解决。甚至100Mb协议需要非常复杂

的编码方法，确保信号频谱在电缆带宽内。

为解决这些任务，千兆以太网开发者必须像快速以太网那样，在其物理层和MAC层做些改变。

13.3.3 保证200m直径的网络

为了在半双工模式下将千兆以太网的最大直径扩展到200m，开发者采取自然步骤。这种解决方案是基于传输最小长度帧所需的时间和PDV之间的比率（第12章中的“案例学习”有介绍）。

最小帧长度从64字节增加到512字节或者4 096bt（不考虑前同步码）。相应地，PDV也增加到4 095bt，假定使用中继器，那么允许网络直径达到大约200m。

将帧长度增加到要求的值，网络适配器使用**扩展**（extension）字段将数据字段填充到448字节，扩展字段是充满零的字段。形式上，帧的最小长度没有发生变化，它保持512位的64字节。这是因为*扩展*（*extension*）字段放在*FCS*字段后面。相应地，这个字段值不包括在效验和之内，并且在长度字段中指定数据字段长度时不需考虑它。扩展字段用于填补载波字段，以便正确地检测冲突。

为减少使用长帧传送短消息的浪费，标准开发者允许端节点依次传送多个帧，不需要通过介质传递给其他站点。这种操作模式称为**突发模式**（burst mode）。假设帧的总长度不超过65 536位或8 192字节，站点可以依次将它们传送。如果站点需要传送多个小帧，可以不用填加*扩展*（*extension*）字段将首个帧填充至512字节。这样，站点可以顺序发送多个帧，直到达到8 192字节的限制（这个限制包括所有帧字节，包括前同步码、头部、数据和校验和）。8 192字节称为*突发长度*（*Burstlength*）。如果站点已经开始传输帧，那么在帧传输过程中达到串长度，站点允许完成这个帧的传输。

“复合”帧的长度的增加延迟了其他站点访问介质；但是，在1 000Mb/s，这个延时并不重要。

13.3.4 802.3z物理介质规范

802.3z标准定义了下列形式的物理介质：

- 单模光纤电缆
- 多模62.5/125光纤电缆
- 多模50/125光纤电缆
- 屏蔽平衡铜电缆

为在传统多模光纤电缆传送数据，标准定义了发射器工作的两种波长：1 300nm和850nm。使用850nm波长的LED的原因是：它们比工作在1 300nm的LED便宜。但是，低的价格对于非常昂贵的技术，如千兆以太网来说是非常重要的。

关于多模光纤，802.3z标准定义了下列规范：1000Base-SX和1000Base-LX。

第一种情况，波长是850nm（S代表短波长）；第二种情况，它是1 300nm（L代表长波长）。对于1000Base-LX，光源是波长为1 300nm的半导体LED。

1000Base-LX允许使用多模（最大段长度高达500m）和单模光纤（最大段长度依赖于发送端功率和电缆质量，能达到几十公里）。

1000Base-CX规范使用屏蔽平衡铜电缆作为传输介质。这种电缆的阻抗为150欧姆。最大段长度仅为25m，因此这种方案仅适合在单个房间内配置。

13.3.5 基于5类双绞线的千兆以太网

每对5类双绞线可以确保100MHz的带宽。为使用这种电缆以1 000Mb/s的速率发送数据，必须使用四对电缆的平行传输数据。

每对的数据传输率可以减少到250Mb/s。但是，就算在这个速率，也必须创造一种编码方法能确保频谱不超过100MHz。例如，4B/5B代码不能解决这个问题，因为在这个速率，155MHz的频率

是信号频谱的主要组成部分。应该记住，每种新技术必须支持本章介绍的半双工模式，还应支持第15章详细介绍的全双工模式。首先，可以看到，同时使用四对线剥夺了网络执行全双工模式的可能性，因为没有保留空闲对用于点到点的同步双向数据传输。

尽管如此，802.3ab任务组找到了这些问题的答案。

对于数据编码，PAM5 代码使用五元电压（−2，−1，0，+1和+2）。因此，每对在每个时钟传送2.322位的信息（$\log_2 5$）。由此，为了达到250Mb/s，250MHz的时钟频率可以减少到2.322次。在这个时钟频率，PAM5的频谱小于100MHz，这意味着可以使用5类电缆不失真地传送。

每个时钟期间，发送8位的信息（而不是$2.322 \times 4=9.288$）。得到总速率是1 000Mb/s。每个时钟传送精确的8位是因为组成有效625（$5^4=625$）位的PAM5代码中只使用256（$2^8=256$）位。接收端使用剩余的组合位控制收到的信息并且从背景噪声中分离合法的组合位。

为组成全双工模式，802.3ab规范的开发者应用从聚合体中获得接收信号的技术。两个发送端执行是通过使用四对中的每对在相同的频段反方向互相发送信息（图13-7）。混合退耦H的设计可以使相同站点的接收端和发送端同时使用双绞线进行传输和接收（类似以太网中的基于同轴电缆的收发器）。

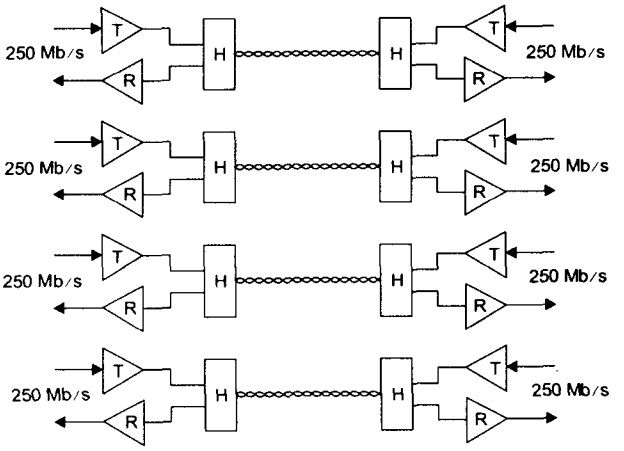


图13-7　使用四对5类UTP的双向数据传输

为从正在传送信息的节点中分离出收到的信息，接收端在混合信号中抽取它的信息。通常，这不是简单的操作，需要特殊的数字信号处理器（DSP）。

小结

- 20世纪90年代早期，需要高速廉价的技术将功能强大的站点连接到网络，这就导致创建了最初的工作组并开发出了一种新技术，这种新技术和以太网一样简单有效但却运行在100Mb/s。
- 专家分为两组，最后在1995年秋天采纳了开发的两种标准：IEEE 802.3委员会通过了快速以太网标准，几乎是100Mb/s以太网的复制，特别创建的802.12委员会采纳了100VG-AnyLAN，它保留了以太网帧格式同时改变了访问方法。
- 快速以太网保留了CSMA/CD，毫无改变的留下了它的算法和位间隔的时间参数（尽管位间隔本身减少了90%）。快速以太网和经典以太网的所有差异显现在物理层。
- 快速以太网标准定义了三种物理层规范：100Base-TX、100Base-FX和100Base-T4。
- 快速以太网的最大直径大概为200m；精确的值依赖于物理介质的特性。快速以太网的冲突域中，不能多于一个1类中继器，或不多于两个2类中继器。

- 基于双绞线的快速以太网允许两个端口通过执行自动协商选择最有效的操作模式。端口可以选择10Mb/s或100Mb/s，同样可以选择半双工或全双工模式。
- 100VG-AnyLAN中，支持需求优先的集中器充当仲裁器的角色，并决定提供站点访问共享介质。
- 千兆以太网给以太网系列的速率层次添加了新的一层：1 000Mb/s。这个层次可以有效地构建有服务器的大规模LAN，并且当千兆以太网主干连接运行100Mb/s的低速率主干时，确保有效的带宽预留。
- 千兆以太网保留了从以太网到快速以太网的许多连续性。千兆以太网使用与以前以太网版本一样的帧格式，并且支持全双工和半双工模式，CSMA/CD仅有稍许改变。
- 特别的802.3ab任务组开发了使用5类UTP的千兆以太网。下列方式确保数据传输率为1 000Mb/s：
  - 通过四对UTP同时传输数据。
  - PAM-5编码，确保使用单个双绞线的数据传输率是250Mb/s。
  - 以全双工模式同步双向传输数据，使用特别的DSP将收到的信号从公共信号中分开。

复习题

1. 需求优先排除的CSMA/CD的什么缺点？
2. 为什么快速以太网的开发者决定保留CSMA/CD？
3. 基于共享介质的快速以太网支持哪些拓扑？
4. 快速以太网的最大直径是多少？
5. 100Base-T4中使用几对电缆传输数据？
6. Ⅰ类和Ⅱ类快速以太网中继器有何不同？
7. 为什么快速以太网仅允许Ⅰ类中继器？
8. 千兆以太网的分组间距的最小值是多少？
9. 由于带宽增加，千兆以太网的开发者必须将最小长度帧增加到512字节。当发送的数据没有填满数据字段时，通过填补使其到达所要求的长度，填补不携带任何信息。千兆以太网采取哪些步骤减少传输短帧的浪费？
10. 千兆以太网采取哪些步骤确保用双绞线使数据传输达到1 000Mb/s？
    A. 提高双绞线电缆的质量
    B. 使用四对电缆代替两对
    C. 增加信号代码状态数量
    D. 执行正交幅度调制
11. 为什么千兆以太网使用多模和单模光纤？

练习题

任务：使用表13-3和表13-4，确定一个配备Ⅰ类中继器的快速以太网具有什么稳定性预留。

提示 当确定快速以太网的正确性时，不使用一或二集线器规则，可以计算PDV值，如第12章Transmash网络“案例学习”中所示。

至于10Mb/s以太网，快速以太网标准提供源数据计算信号PDV。但是，这个数据形式和计算方法已经改变。快速以太网提供的数据是关于每个网段引入的双倍延时，不需要将网络划分成左、中和右。另外，网络适配器引入的延时考虑了前同步码。因此，PDV必须与512bt比较(即，传输没有前同步码的最小长度帧的时间)。

对于I类中继器，RTT的计算如下：使用电缆传输信号引入的延时按照表3-3提供的数据计算，考虑信号必须经过电缆两次。两个网络适配器（或交换机的端口）和中继器的相互作用引入的延时如表13-4所示。

考虑I类中继器引入的双倍延时是140bt，可以对任何网络配置计算RTT，（表13-2提供，考虑最大长度的电缆段）。如果结果值小于512，那么按照冲突检测算法，这个网络配置是正确的。802.3标准建议为网络运行平稳保留4bt。但是，允许在0到5bt间选一个值。

表13-3　电缆引入的延时

电缆类型	双倍延时 （比特间隔每米）	最大长度电缆的 双倍延时
3类UTP	1.14bt	114bt(100m)
4类UTP	1.14bt	114bt(100m)
5类UTP	1.112bt	111.2bt(100m)
STP	1.112bt	111.2bt(100m)
光纤	1.0bt	412bt(412m)

表13-4　网络适配器引入的延时

网络适配器类型	环路的最大延时
两个TX/FX适配器	100bt
两个T4适配器	138bt
一个TX/FX适配器和一个T4适配器	127bt

第14章 共享介质的LAN

14.1 引言

本章考虑以太网以外的几个共享介质LAN技术。成员包括令牌环和FDDI，它们成功地用于LAN已有很长时间了，这些LAN需要较高的性能和可靠性以及广泛的覆盖范围。交换LAN到来之前，这些技术在以上几个方面已经超过了以太网。所以，建立LAN主干网和为金融和政府机构创建网络，也就是说当网络的性能和可靠性非常重要时优先考虑这些技术。令牌环和FDDI采用确定性访问方法，可以更有效地共享传输介质，并且提供支持QoS的实时业务。

令牌环和FDDI采用环形拓扑的物理链路，能自动控制网络的运转。FDDI网络另外还能确保故障后自动恢复网络。为提供这种服务，它们使用双环连接节点；这方面，类似于SDH网络。

无线通信介质通过物理设备共享。本章包含两种无线通信技术，IEEE 802.11和蓝牙（IEEE 802.15.1）。前者可以创建无线LAN;后者属于个人区域网（PAN）。每种技术都有其访问介质的方法。

14.2 令牌环

令牌环（Token Ring）技术是IBM公司1984年开发的，并送到IEEE 802委员会作为建议的标准项目。IEEE 802委员会把这项技术作为1985年采用的802.5标准的基础。长期以来，IBM公司建立了基于不同类别计算机的LAN，从大型机和强大的小型机到个人电脑，都使用令牌环作为主要的网络技术。但是，最近几年，以太网系列已经占据统治地位，甚至在IBM的产品中。

令牌环网执行两种位速率：4Mb/s和16Mb/s。单个环中的工作站不允许执行不同的速率。令牌环网执行16Mb/s，包括对用4Mb/s网络中的标准访问算法的一些改进。

令牌环比以太网复杂。它能提供一些基本的容错特性。令牌环网定义了控制网络操作的特殊程序，这些程序使用环拓扑固有的反馈特性：发送的帧总会回到发送端。某些情况下，网络故障会自动纠正。例如，丢失的令牌可以自动恢复。在另一些情况下，网络仅能报告发现的错误，并支持人们手工消除它们。

为控制网络操作，一个工作站代表**有源监视器**（active monitor）的角色。有源监视器在环初始化过程中选定。最大介质访问控制（MAC）地址值是判断标准。如果有源监视器失效，重复环的初始化过程，并选择新的有源监视器。为确保网络能够检测到有源监视器的失效，后者产生一个特殊的帧通知其他工作站点它的出现。每3秒发出一次（假定有源监视器正常工作）。如果7秒以后还没发出这样的帧，其他工作站点就启动寻找新的有源监视器的程序。

14.2.1 令牌传递访问

令牌环网是基于令牌传递规则访问共享介质，第12章描述MAC层功能时介绍过。我们将详细地介绍这种方法的一些特性，即802.5标准描述的**令牌环4Mb/s（Token Ring 4Mb/s）**技术的典型性质。

令牌环网中，每个站点都只直接从一个站点接收数据：环中的前一个站点。每个站点把数据发送到与它相邻的最近下游站点。

收到令牌后，站点对它进行分析。如果此站点没有数据发送，就将令牌传到下一个站点。当令牌被传到有数据要发送的站点时，它把令牌收取，获得访问物理介质的权利并发送它的数据。

此后，站点陆续向环发送特殊格式的帧。这些帧包含源地址和目的地址。

被传送的数据总是逐站地沿环的一个方向传播。环中所有的站把帧一位一位的转发出去，起了中继器的作用。如果帧到达目的节点，这个站点识别出它的地址，把帧复制到内部缓存器，并在帧中插入接收确认指示符。将数据帧发送到环的站点，再次收到包含确认已经接收到信号的帧，从环中收取帧并发送新的令牌到网络，允许其他站点发送数据。

图14-1提供了时间图表解释这里描述的介质访问机制。示例中在有六个站点的环中，将分组A由站点1发送到站点3，这个分组中设置了两个标记：地址识别标记A和指示符C，表明分组已经被复制到内部缓存器。（图中，在分组内用星号表示）当分组返回站点1，发送端通过源地址识别出它的分组并从环中收取分组。站点3设置标记通知发送端分组已被成功地发送到目的节点并复制到内部缓存器。

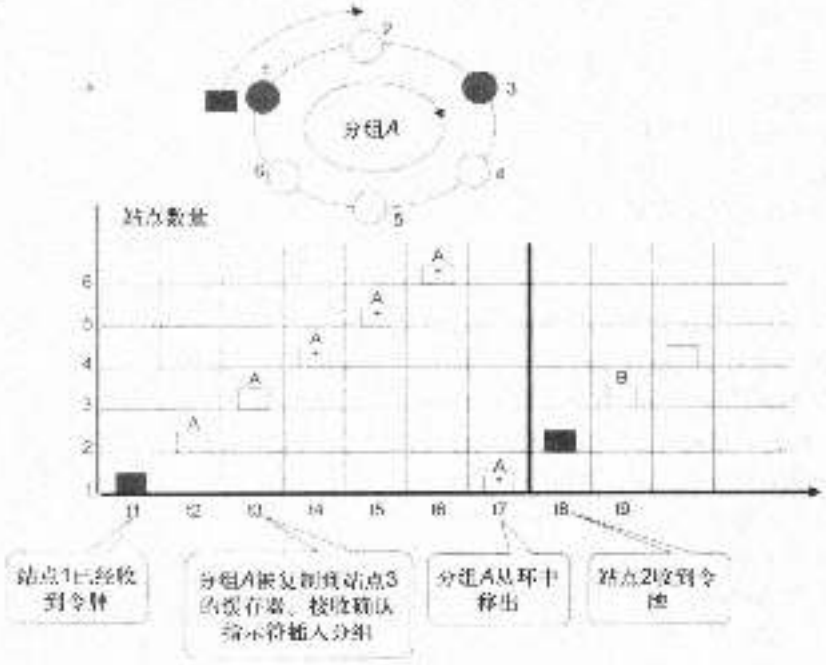


图14-1　令牌传递访问

令牌环网络中共享介质独占的时间限制为固定值，称为**令牌保持时间**（token-holding time）。当这个时间消逝，站点必须停止发送数据（允许完成当前帧的传输）并沿环发送令牌。令牌保持期间，站点可以发送一个或多个帧，依赖于帧的大小和令牌保持间隔的持续时间。默认情况下，令牌保持时间设置为10ms，802.5标准没有严格定义最大帧长。按规定，4Mb/s网络，帧长是4KB；16Mb/s网络，帧长是16KB。选择这些值是因为令牌保持时间内，站点必须至少传送1帧，4Mb/s网络，10ms可以传送5 000字节；16Mb/s网络，相同时间内可以传送20 000字节。最大帧长度的选择应当有所保留。

令牌环16Mb/s（Token Ring 16Mb/s）使用稍微不同的环访问算法，称为**令牌提前释放**（early token release）。按照这种算法，邻居当发送完帧的最后位时站点将访问令牌立即发送到离它最近的下游，不需要等待帧设置接收确认位A和C。这样，环的带宽得到更有效地利用，因为多个站点的帧同时延环移动。然而，只有一个站点能够产生帧：拥有访问令牌的站点；所有其他站点此时只能转发其他节点发送过来的帧，以便保持时分共享原则。这样就加快了令牌传递过程。

对不同类型的消息，传送的帧被分配不同的优先级，从0（最低）到7（最高）。传送站点决定

帧的优先级。令牌环协议使用高层协议的服务接口接收这个参数，例如应用层协议，令牌总有具体的当前优先级。站点有权获得传送给它的令牌，仅当要发送帧的优先级等于或高于令牌的优先级时。否则，站点必须沿着环将令牌发送到下一个站点。

有源监视器负责网络中单个令牌副本的存在。如果有源监视器长时间（如2.6s）没有收到令牌，那么它就产生一个新的令牌。

令牌环优先级访问的目的是为应用程序支持QoS要求。但是，应用程序的开发者希望LAN不使用这种性能。

14.2.2 令牌环物理层

IBM公司开发的令牌环标准最初是为使用**多站点访问部件**（multistation access unit，MAU或MSAU）集中器（即多站点访问设备）（图14-2）建立网络星路做准备。令牌环网络可以包含260个节点。集中器的使用赋给令牌环网络物理"星形"拓扑，它们的逻辑拓扑是环。

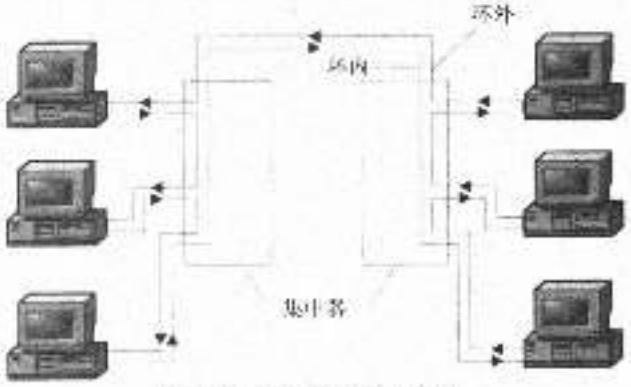


图14-2 令牌环网的物理结构

令牌环集中器可以是有源或无源。**无源集中器**（passive concentrator）简单地通过内部旁路连接端口，以便将连接这些端口的站点形成环状，无源MSAU集中器不放大和再同步信号。这样的设备可以看做是简单的交叉设备，第一例外：如果与计算机连接的端口停止运转，MSAU确保信号从特定端口旁路通过，这个要求是为确保环路的连接性独立于连接的计算机状态。通常，旁路端口使用继电器电路，由网络适配器的直流供电。当网络适配器停止运转，平常关闭的继电器接触器连接输入端到它的输出端。

有源集中器（active concentrator）执行信号再生功能，因此，有时称为中继器。

如果集中器是无源器件，那么如何确保包含几百台计算机的大范围网络，长距离高质量的信号传输？答案很简单，在这种情况，每个网络适配器承担信号放大器的角色；同步部件的角色托付给环中有源监视器的网络适配器。每个令牌环网络适配器有一个中继器部件能够再生和同步信号。但后者的角色仅由有源监视器的中继器部件承担。

通常，令牌环网是复合的星形——环形结构，端节点按照星形拓扑连接MSAU，MSAU集中器使用特殊的环内和环外端口互联形成一个物理的主干环。

令牌环允许使用各种类型的电缆连接端节点和集中器：1类STP，3类STP，6类STP（IBM电缆系统类型）和光纤电缆。

当使用来自IBM电缆系统类型的1类STP时，可以连接260个站点到环上，该种电缆的最大长度是100m。如果使用非屏蔽双绞线，最大工作站点的数量减少到72个且电缆长度减少到45m。

使用1类STP电缆时，无源MSAU集中器间的距离可以到达100m，使用3类UTP，这个距离减少到45m。有源MSAU集中器间的最大距离依电缆类型是730m或365m。

令牌环网最大环长度是4 000m。

说明　令牌环网中的最大环长度和环中工作站点的数量的限制没有以太网中的限制那样严格。令牌环中，限制涉及令牌沿环周转的时间，尽管对限制选择的定义有另外的考虑。假定环包含260个工作站点。令牌保持时间是10ms，最坏的情况下，令牌在2.6s后回到有源监视器。这个时间等于令牌周转的超时时间。重要的是，令牌环网中，网络节点的网络适配器的所有超时值都可以调整。因此，就能建立有许多站点和较长环的令牌环网络。

14.3　FDDI

FDDI—光纤分布式数据接口（Fiber Distributed Data Interface）——是第一个使用光纤作为传输介质的技术。20世纪80年代开始为LAN使用光纤链路研究新型设备和技术，不久以后，这些链路被用于WAN。1986年到1988年，ANSI的X3T9.5研究组开发了第一个版本的FDDI标准，确保使用100Km长的双线光纤环传输帧的速率是100Mb/s。

14.3.1　主要的FDDI特性

FDDI在许多方面基于令牌环，发展和改进了它的主要思想。FDDI研发者的主要目的：

- 提高数据传输的位速率到100Mb/s。
- 通过在故障后引入恢复的标准进程提高网络的容错能力，例如电缆故障、节点的错误操作、集中器故障和链路高电平噪声。
- 确保同步和异步（时延-敏感）业务最大限度地使用潜在的网络带宽。

FDDI建立在两个光纤环基础上，形成网络节点间数据传输的主要和保护路径。使用两个环是提高FDDI网络容错的主要方法。

要从提高的容错潜能中受益的重要节点必须与两个环相连。为了能使用光纤发送数据，FDDI执行4B/5B逻辑编码与不归零反转（NRZI）编码。这种传送信号的方法使用的通信链路时钟频率是124MHz。

在正常操作模式下，发送的数据只通过**主要环**（primary ring）的所有节点和所有部分。这个模式称为**通过模式**（thru mode）（即运输模式）。这种模式不使用**次要环**（secondary ring）。

如果故障发生或主要环的一部分不能传输数据（由于电缆损坏和节点故障引起），主要环连接到次要环（图14-3），再次形成闭合环。这种网络操作模式称为**倒换模式**（wrap mode）。FDDI的集中器或网络适配器，或两者执行倒转操作。为简化过程，沿主要环的数据传输是单向的（图中是逆时针方向）。沿次要环的数据传输是反方向的（逆时针方向）。因此，当两个环形成一个普通环时，站点发送端仍然连接邻近站点的接收端，可以正确接收和传输数据。

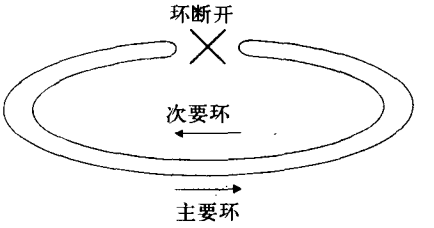


图14-3　故障后重新配置FDDI环

FDDI标准注重的过程是可以检测网络设备的故障并且按要求执行重新配置。FDDI通过配置数据传输路径完善令牌环的错误检测机制，这些路径是基于次要环所确定的保留链路。

FDDI网络能够在个别元件发生故障后恢复其可用性。在多种故障情况下，网络被分解为几个独立的彼此互不相连的网络。

FDDI中的环被认为是数据传输的公共共享介质，因此开发了一种特殊的访问方法。这种方法

类似令牌环网的访问方法；又称为**令牌传递方法**（toke-passing method）。FDDI站点使用早期的令牌释放机制，类似于令牌环16Mb/s网络使用的机制。

令牌环和FDDI方法的不同之处：

- 与令牌环不同，FDDI网络的令牌保持时间并不固定。相反，这个时间依赖于环的负载。当负载低时，令牌保持时间增加；拥塞期间，可能降到零。访问方法中的这些变化只涉及异步业务，对小的帧传输延时不敏感。同步业务中，令牌保持时间保持固定。
- FDDI不执行与令牌环类似的帧优先级机制。研发者认为把业务分为八个级别是多余的，将所有业务分成两个类别即同步和异步就足够了。当网络过载时，后种划分仍可工作。其他方面，在MAC层的环工作站点之间发送帧与令牌环相同。

图14-4说明了FDDI协议栈与七层OSI模型的对应关系。FDDI定义了物理层协议和数据链路层的MAC子层协议。正如许多其他LAN技术，FDDI使用IEEE 802.2标准定义的逻辑链路控制（LLC）协议。

FDDI技术的出色性质是**站点管理**（station management，SMT）层。这层执行的功能涉及管理和监视FDDI协议栈的所有其他层。FDDI网络的每个节点参与控制环。因此，所有节点转发特殊SMT帧以便控制网络。

其他层的协议也参与确保FDDI网络的容错。例如，物理层消除物理原因导致的网络故障，如电缆中断。另一方面，MAC层帮助弥补逻辑的网络故障，如丢失集中器端口间传送令牌和帧所必需的内部路径。

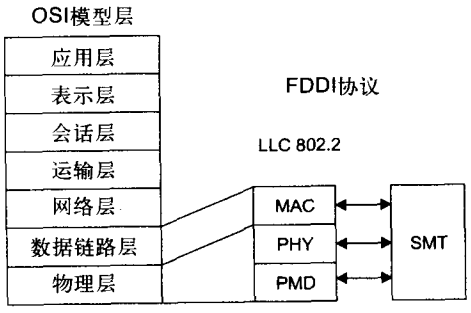


图14-4 FDDI协议栈

14.3.2 FDDI容错

为保证容错，FDDI标准提供两个光纤环：主要和次要。

FDDI定义了两种类型的站节点：**站点**（station）和**集中器**（concentrator）。为将站点和集中器连接到网络，可以使用如下方法：

- **双连接**（Dual attachment，DA）——同时连接主要环和次要环。使用这种方法连接的站点和集中器分别称为**双连接站点**（dual attachment station，DAS）和**双连接集中器**（Dual attachment concentrator，DAC）。
- **单连接**（single attachment，SA）——仅连接主要环。使用这种方法连接的站点和集中器分别称为**单连接站点**（single attachment station，SAS）和**单连接集中器**（Single attachment concentrator，SAC）。

通常，尽管不是必须的，集中器是DA，站点是SA，如图14-5所示。为了简化连接网络的设备，它们的集中器被标记。设备中*A*和*B*类的集中器必须双连接；主（M）集中器必须在集中器中作为该的单连接，必须有一个S类的从响应连接器。

当双连接的设备间的单个电缆断开，FDDI网络能够对集中器端口间的内部帧传输路径自动重新配置，这样FDDI网络就能继续正常运行（图14-6）。

双电缆断开将产生两个独立的FDDI网络。当单连接的站点的电缆断开时，站点从网络中隔离出来，但环通过集中器内部路径的重新配置能继续运转：连接站点的端口M将从公共路径中删除。

为了保证当双连接的站点停止工作后网络能继续运转，这些站点必须配备**光纤旁路开关**（optical bypass switch），电源中断发生时为光束创建一条旁路。

最后，DAS 或DAC连接一个或两个集中器的两个主端口，建立了主和预留链路的树状结构。

默认设置情况下，端口*B*支持主链路，端口*A*支持保留链路。这种配置就是**双归宿**（dual homing）。

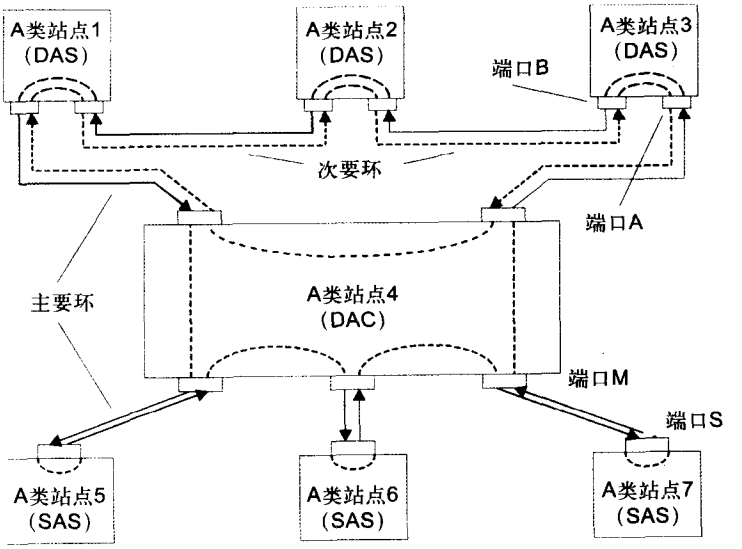


图14-5　FDDI环连接的节点

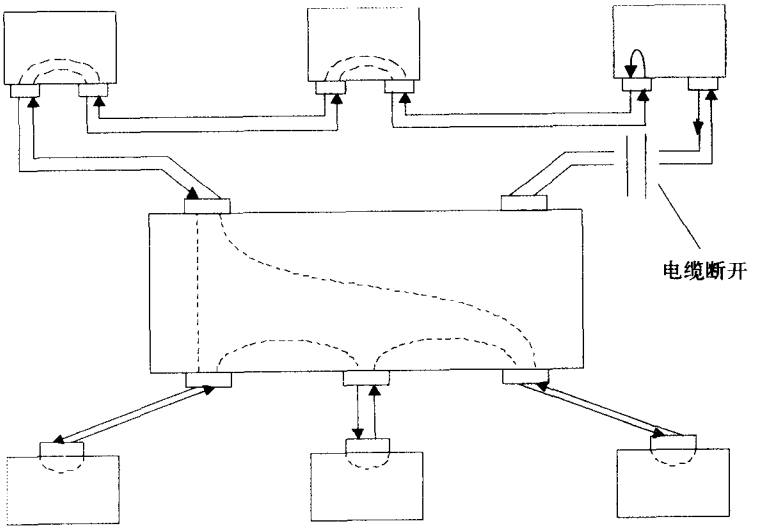


图14-6　电缆断开后FDDI网络重新配置

通过不断追踪令牌环和帧的流通时间间隔及检测网络中相邻端口的物理连接是否存在，工作站和集中器的SMT层可以支持容错。FDDI网络没有专门的有源监视器。所有站点和集中器都有相同的权利，它们中的任何一个都能对网络进行再初始化，如果检测到违背正常操作还能对其重新配置。

特殊光交换机能够改变光束方向并具有复杂的结构，它对网络适配器和集中器内部路径进行重新配置。

FDDI环总长度的最大值是100km；环中双连接的站点数量的最大值是500。

开发FDDI是为了将其用于重要的网段——连接大型网络的主干，例如，建立网络——并连接高性能服务器到网络。因此，其最重要的目的包括确保网络节点间长距离高速数据传输和协议层

次的容错。所有这些目标都已实现。由此，FDDI保证高质量但费用昂贵。甚至使用基于双绞线的便宜方案也不能明显减少单个节点连接到FDDI网络的费用。因此，FDDI主要应用领域是连接多个建筑物的主干网和大城市的MAN。

14.4 无线LAN

14.4.1 无线LAN的特殊性质

无线LAN被认为是有线LAN的补充而不是竞争对手。但是，无线LAN并不总被这样认为。20世纪90年代中期，另一种观点非常流行：预测随着时间的过去，更多的LAN将转到无线技术。无线LAN的优势是明显的：它们更容易配置和升级，不需要大量的电缆基础设施。确保用户移动性是另一个优势。但是，无线网络使用不固定和不可预测的无线介质会产生许多问题。第8章讨论了信号在这种介质中传播的特殊性质。

来自各种家庭装置的外部噪声（External noise）、其他通信系统、大气噪声和信号反射都会对信号的可靠接收产生巨大冲击。LAN主要目的是连接建筑物内的计算机，建筑物内无线电信号的传输比户外复杂。IEEE 802.11标准提供"有标准金属桌子和打开的门的简单方形房间"的信号强度分布图（图14-7）。标准强调这种分布是静态的；实际上，模型是动态变化的。对比，房间内各种物体的移动会明显改变信号的分布。

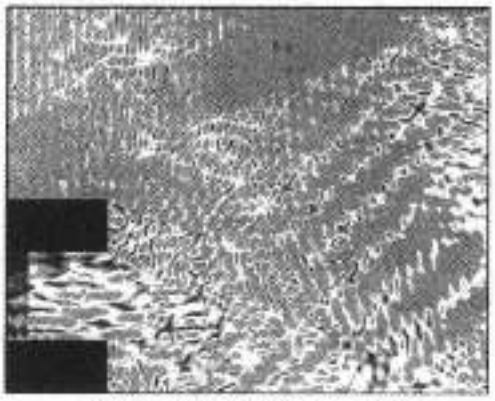


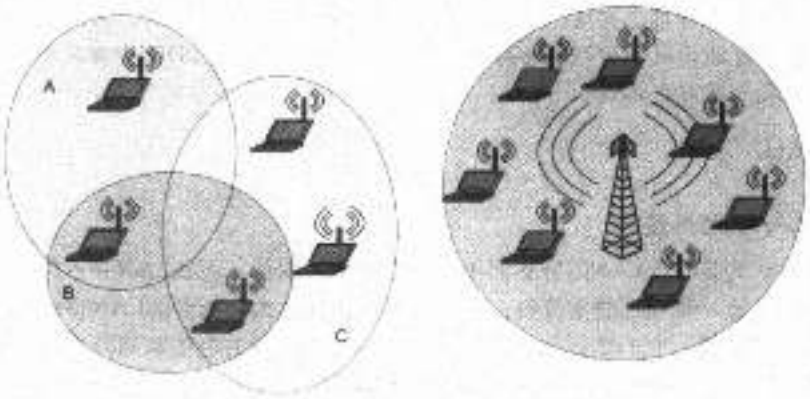
图14-7 无线电信号的强度分布

*扩频（spread spectrum）*方法可以使噪声对有用信号的影响减少。除此之外，无线网络广泛使用*前向纠错（forward error correction，FEC）*方法和确保丢失帧重传的协议。然而，实践证明没有什么能阻止单位使用有线LAN，尽管不可能不使用电缆系统，但大多数单位更喜欢使用这样的网络。

不均匀的信号强度分布不仅会导致正在传输的信息中出现位错误，而且还会产生无线*LAN覆盖区域的不确定性（uncertainty of the coverage zone of a wireless LAN）*。有线LAN没有这样的问题，因为连接建筑物和校园电缆系统的所有设备都接收信号并参与LAN运行。无线LAN没有精准确定的覆盖区域；通常采用的描述区域的符号，如圆和六边形，都是抽象概念。实际上，在这些规则的覆盖区域的某些部分，信号很弱以至于这些区域内的设备无法接收和发送信息。

图14-7所示的图形很好地解释了这种情况。有必要强调随着时间的过去，信号分布形状会发生相当大的改变，LAN的结构也相应地发生变化。所以，就算固定的网络节点（认为是不移动的）也必须考虑不能完全连接无线LAN。假定信号是全向的，无线电信号按离源端距离平方的比例衰减，防止产生全连接的拓扑。因此，没有基站，网络节点的某些节点对不能通信因为它们处在其对方发送端覆盖区域的外面。

图14-8所示的例子说明了LAN片段。无线网络缺少全连通性导致**隐藏终端问题**（hidden terminal problem）。当两个节点都处在彼此限度之外时，这种问题就会出现（图中的节点*A*和*C*）。但是，第三个节点*B*能够同时收到*A*和*C*的信号。假定无线电网络使用传统的基于载波侦听的访问方法，如CSMA/CD。在这种情况，冲突比传统有线网络更频繁。例如，假定节点*B*正在与节点*A*交换信息，节点*C*检测到介质空闲，就开始传输它的帧。结果，节点*B*附近的信号失真（即冲突发生）。有线LAN发生这类冲突的概率更低。



a）Ad hoc无线网络　　b）有基站的无线网络

图14-8　无线LAN的连接

无线电网络的**冲突检测**（Collision detection）也很复杂。因为节点本身的发送端的信号抑制了远处发送端的信号。所以，检测出信号失真很困难。

无线网络的访问方法抛弃了载波侦听和冲突检测。它们使用各种**冲突避免**（collision avoidance, CA）方法，包括**轮询**（polling）。

基站（base station）的使用可以提高网络连通性（图14-8b）。这些基站通常有较高的功率，配置的天线能够更加均匀和容易地覆盖要求的区域。因此，无线LAN的所有节点都能与基站交换信息，基站充当网络端节点之间交换信息的转接节点。

无线LAN被认为在非常困难或不可能使用有线LAN的应用方面大有前途。无线LAN的主要应用领域是：

- 可选择供应者的小区访问（*Residentiaal access*），没有有线访问住在公寓内的客户。
- 机场、火车站等的移动访问（*Mobile access*）。
- 在不可能安装现代电缆系统的建筑物内组建LAN，如原始内貌的历史建筑。
- 需要临时LAN，例如会议。会议参与者出席会议时不能使用有线连接。
- *LAN的扩展*（*LAN Extension*）。例如，公司的一个建筑物，如厂房或测试实验室可能与其他建筑物隔离。建筑物中的极少工作场所使安装电缆效率低。因此，无线通信证明更合理。
- *移动LAN*（*Mobile LAN*）。当用户从房间到房间或者从建筑物到建筑物移动时候，访问LAN。此时无线网络没有竞争者。这种用户的典型例子是，医生看望病人并使用笔记本连接医院的数据库。

目前，移动LAN真正完全覆盖大片区域，如移动蜂窝电话网络那样。然而它们有潜能这样做。为数据传输建立的移动蜂窝网络领域内，无线LAN技术必须与**第三代**（3G）**移动蜂窝网络**（third-generation（3G）mobile cellular network）竞争。2G**移动蜂窝网络**（2G mobile cellular network）不是主要的竞争者，因为它们主要是为语音传输而开发的。它们在数据传输领域的能力局限在几千位每秒的速率，无线LAN确保十几兆位每秒的速率。但是，3G系统的传输速率期望在144Kb/s到2Mb/s之间（距离基站近的地方能达到后面的速率）。这样，竞争就非常激烈。

这章的后面，将讨论最流行的无线LAN标准——IEEE 802.11。除IEEE 802.11之外，这个领域还有其他标准，特别是，欧洲电信标准机构（ETSI）开发的HIPER-LAN1标准。然而，大多数制造商按IEEE 802.11规范生产设备。

14.4.2 IEEE 802.11协议栈

大体上，这个标准的协议栈对应于802委员会标准的公共结构。这意味着包含物理层和运行在LLC层之上的MAC层。如同所有802系列的技术一样，802.11定义了两个最底层：物理层和MAC子层。LLC子层（LLC2）执行它自身的功能，这些功能对所有以太网技术都是标准的。因为无线介质比导向介质容易产生帧的失真，所以LLC更可能使用LLC2模式。这不依赖于802.11技术，因为LLC操作模式是由高层协议选择的。

IEEE 802.11协议栈的结构如图14-9所示。

物理层（*physical layer*）按照不同的频率范围，编码方法和信息速率有不同的规范。物理层的所有变量共用相同的MAC层算法。但是，MAC层的一些时间参数依赖使用的物理层。

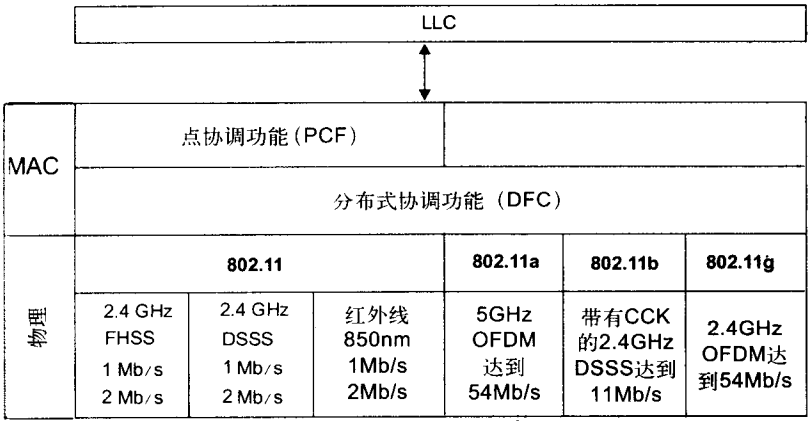


图14-9 IEEE 802.11协议栈

1997年，802.11委员会采纳了这个标准，规定MAC层功能支持物理层的*三种不同形式确保数据传输率*（*three variants of the physical layer*）是1Mb/s和2Mb/s。

- 第一种形式是使用850nm的**红外波**（infrared wave）作为传输介质。它们由半导体激光二极管或者发光二极管（LED）产生。因为红外波不能穿透墙壁，所以这种LAN被限制在视线不被阻挡的区域范围内。标准规定了三种不同的波传播：全向天线、天花板反射和集中方向发射。第一种情况，用透镜系统把窄带光束发散。后一种形式目的是在两个建筑物间组织“点对点”通信。
- 第二种形式是使用ITU 推荐的2.4GHz**微波段**（microwave range），它在大多数国家都未获得许可。一种物理层的微波形式是跳频扩频（FHSS），另一种是直接序列扩频（DSSS）[⊖]。FHSS的每个窄带信道宽度是1MHz。使用两种信号状态（频率）的FSK调制得到的速率是1Mb/s，使用四种信号状态可提供2Mb/s的速率。当采用FHSS时，网络可以含有*蜂窝*；为消除相互干扰，相邻*蜂窝*使用正交频率序列。信道数量和信道间的转换频率都有规定，因此无线LAN的设计者应当考虑某个具体国家频谱规则的特殊性质。例如，美国2.4GHz的频段可有79个信道，每个信道花费的最大时间不超过400 ms。
- 第三种形式是使用基于DSSS编码的2.4GHz的*微波段*（*microwave range*）。DSSS编码使用11位码10110111000作为片序列。每位都用BPSK（1Mb/s）或QPSK（2Mb/s）编码。1999年，允许两种其他形式的物理层：802.11a和802.11b。
- *802.11a*规范通过使用更高的5GHz频率范围提高了信息速率。为此目的，它用了这个范围的300MHz，正交频分复用（OFDM）和前向纠错（FEC），它能使用的信息速率包括6、9、12、

⊖ 这种方法的更多细节信息在第10章介绍。

18、24、36、48和54Mb/s。802.11a的5GHz频段很少使用并能确保高的信息速率。但是，使用这个频段会产生两个问题：第一，在这个频率工作的设备太昂贵。第二，在一些国家，这个频段已被使用。

- 第二个规范，IEEE 802.11b仍然使用2.4GHz，可以使用便宜的设备。为了达到可与经典以太网相比的11Mb/s速率，这种技术使用更有效的DSSS方法，采用**补偿代码键控**（complementary code keying，CCK）——改进的调制机制，于1999年替换了无线数字网络的**巴克码**（Barker code）（见第10章）。

最新的802.11工作组标准的物理层，IEEE 802.11g在2003年夏天通过。

- IEEE 802.11g也工作在2.4GHz频段但确保数据传输率达到54Mb/s。这个规范同样使用OFDM。直到现在，美国的法令才允许在2.4GHz频段仅采用扩频技术。此限制的取消促进了新的研究和创新，结果出现了高速无线技术。为提供对802.11b的向后兼容，采用了CCK。

802.11网络的直径依赖于许多参数，包括使用的频段。通常，无线LAN的直径在100m到300m之间。

无线LAN中MAC执行的功能多于有线LAN。

802.11标准的MAC功能包括：

- 支持访问共享介质。
- 当多个基站可用时，确保站点的移动性。
- 保证与有线LAN同样的安全性。

14.4.3　802.11LAN的拓扑

802.11标准支持两种类型的LAN拓扑——Ad-hoc网络，如我们所知的基本服务集（BSS），和具有基础设施的网络称为扩展服务集（ESS）。

按照802.11的术语，*Ad-hoc*网络（*Ad-hoc network*），又被称为**基本服务集**（basic service set，BSS），是由**独立的站点**（individual stations）创建的。它们不含有基站，网络中的节点直接和其他节点通信（图14-10）。要成为BSS的一员，站点必须执行连接程序。

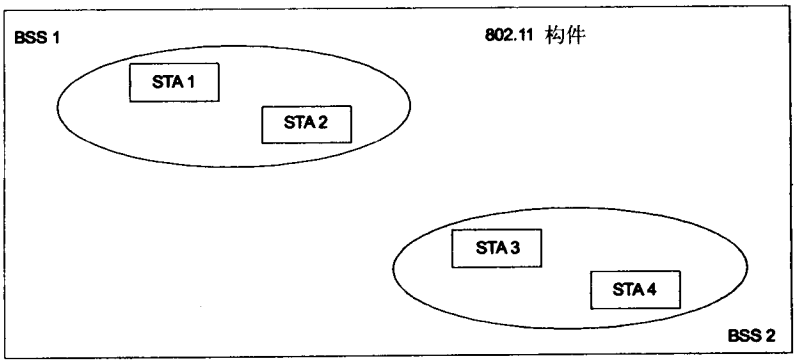


图14-10　基本服务集

BSS并不是传统意义上的覆盖区域的蜂窝，因为它们位于彼此相隔很远的地方。它们还可以部分或全部交叠。802.11在这方面提供了自由的网络构架。

在有基础设施的网络中，一些站点是基站。按802.11的术语，基站称为**访问点**（access points，AP），执行AP功能的站点是BSS的成员（图14-11）。所有网络的AP由分布式系统（*distribution system，DS*）互连。DS的角色可以由连接站点的介质（如，无线电或红外波）或其他介质如电缆担当。有了DS，AP可以执行**分布式系统服务**（distribition system service，DSS）。

DSS的任务是在站点间发送分组，这些站点不能（或不想）直接相互作用。使用DSSS最重要的理由是，如果站点属于不同的BBS，那么这些站点传送它们的帧到其AP，AP再使用DS把这些帧传送到AP，这个AP服务于目的站点所在的BSS。

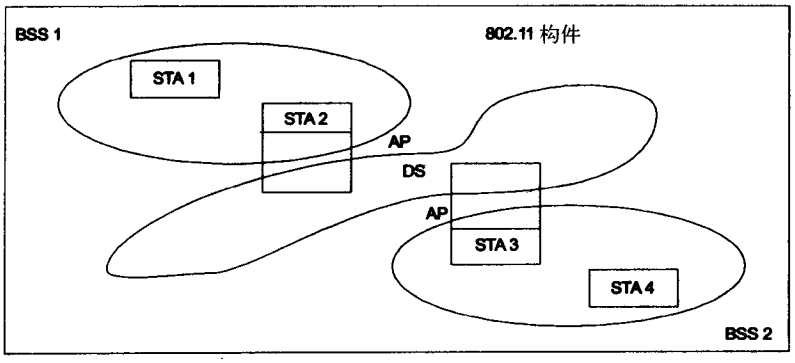


图14-11 分布式系统的扩展服务集

按照802.11术语**扩展服务集**（Extended Service Set，ESS）网络是一个包含多个由DS连接的BSS的网络。

ESS确保站点的移动性，因为它们能从BSS移动到BSS。站点和AP的MAC层功能确保这些移动。ESS也能和有线LAN通信。为此目的，DS必须包含一个**门桥**（portal）[⊖]。

14.4.4 访问共享介质

站点使用共享介质的方法实现下列目的：

- 单个BSS内互相直接传送数据。
- 单个BSS内使用AP作为转发节点传输数据。
- 使用两个AP和DS在BBS间传送数据。
- 使用AP、DS和门桥在BBS和有线LAN间传送数据。

在802.11网络中，MAC层确保访问共享介质的两种模式：

- **分布式协调功能**（distributed coordination function，DCF）
- **点协调功能**（point coordination function，PCF）

1. 分布式协调功能访问模式

首先，考虑使用DCF提供访问的方法。这种方法执行著名的CSMA/CA算法。它属于基于载波侦听的CA算法。同时，它使用“时间槽算法”。这种方法采用间接冲突检测，替代了基于介质状态的直接冲突检测过程，后者在无线网络中的效率很低。每个发送的帧必须经目的站点发送的ACK帧确认。如果在预定时间间隔周期没有收到ACK，那么发送端认为发生了冲突。

使用时间槽访问算法要求站点是同步的。802.11技术中，这个问题已经圆满解决：当下一帧的传输完成时，开始对时间间隔计数（图14-12）。不需要传输特殊的同步信号并且时间槽大小不会限制分组的长度，因为当决定发送帧时才需要考虑时间槽。

需要发送帧的站点首先检测载波。当它记录了帧传输的结束，必须等待帧间间隔（*interframe space*，*IFS*）的时间间隔。如果IFS消失后，介质仍然空闲，那么开始时间槽的计数。每个时间槽都有**时隙时间**（SlotTime）周期。假定介质是空闲的，时间槽启动才能开始帧的传输。站点选择时间槽是基于**退避二进制指数算法**（trunceted binary exponential back-off Algorithm），类似用

⊖ 门桥的功能没有详细定义：它的角色可由交换机或路由器担当。

于CSMA/CD。时间槽数量是一个随机的整数，取值范围是{0，CW}（CW表示竞争窗口(contention window)）。

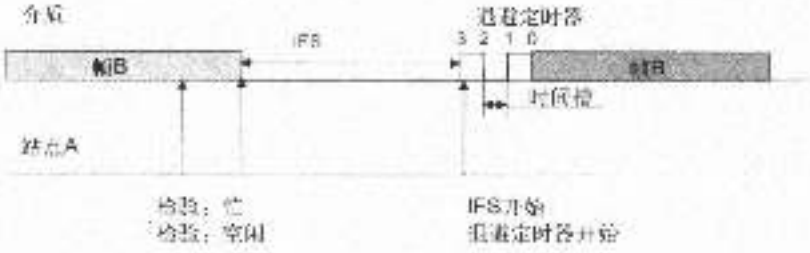


图14-12　DFC算法

本章后面将介绍选择时间槽大小和竞争窗口大小的方法。此时，用一个实际例子来说明这种精妙的访问方法（图14-12）。假定站点A基于退避二进制指数算法选择3作为传输的时间槽。选择时间槽数量后，站点分配3给退避定时器（back-off timer）（其作用将在后续讨论中说明白），并在每个时间槽开始时检测介质。如果介质可用，退避定时器的值减1。如果减到0时，开始传输数据。

因此，算法确保了所有时间槽包括当前那个都是可用的。这是传输开始的基本和必备的条件。

如果介质在某个具体时间槽开始时正好忙，计数器是"0"（即，它不减1）。如果这种情况发生，站点开始新一轮的介质访问，只改变为传输选择时间槽的算法。正如前一轮中，站点继续检测载波。当介质再次空闲时，站点停止一段时间，这段时间等于IFS的持续时间。如果介质在时间间隔消失后仍然空闲，站点使用退避定时器的冻结值作为时间槽数，执行以上描述的检验空闲时间槽的过程并从定时器的冻结值开始减少它的值。

时间槽大小依赖于信号编码的方法，因为对FHSS方法，时间槽的大小是28μs；对DSSS方法是1μs。选择的时间槽大小应该超过任意两个站点间的信号传播时间，加上站点正确识别介质可用性所要求的时间。如果这个要求满足，当检测到时间槽先于所选定传输的那个，每个站点都能正确识别载传输的开始。这样，反过来意味着：

当多个站点选择相同的时间槽传输时，冲突发生。

如果冲突发生，帧损坏，并且没有来自目的节点的ACK帧。如果站点在预定时间周期内没有收到来自目的节点的ACK，记录一个冲突并重传帧。每次试图传输帧失败后，选择的时间槽数量所在范围[0，CW]加倍。例如，初始窗口大小选择8（即，CW=7），第一次冲突后窗口必须设置为16（CW=15），第二次冲突后窗口的大小必须设置为32，等等。802.11标准规定CW的选择必须依赖于无线LAN的物理层类型。

正如CSMA/CD，试图重传一个帧失败的次数是有限制的。但是，802.11标准并不提供精确的上限值，当尝试失败的次数达到上限N，帧被丢弃，且冲突计数器设置为0。通常，如果在尝试失败的数次后，站点能够成功发送帧，计数器也设置为零。

DFC采用特别的方法消除**隐藏终端影响**（hidden termianal effect）。为此目的，站点需要获得介质，并且按照前面描述的算法，决定在特定的时间槽开始传输帧，发送一个短的请求发送帧（*request-to-send frame，RTS*）代替数据帧给目的站点。目的站点回复一个清除发送帧（*clear-to-send frame，CTS*）。此后源站点发送数据帧。CTS帧必须包含关于介质捕获的信息，为在发送站点范围之外但在目的站点覆盖区域内的所有站点能够得到它（即，所有站点对发送端而言是隐藏终端）。

说明　按照802.11标准，最大帧长度是2 346字节，RTS长度是20字节，CTS帧为14字节。因为RTS和CTS帧比数据帧短得多，由RTS或CTS帧冲突引起的丢失比数据帧

冲突引起的丢失要小。交换RTS-CTS帧的过程是可选择的：当网络负载低时，可以不用它，因为冲突很少，这意味着没必要花费附加时间执行RTS-CTS过程。

2. 点协调功能访问模式

如果BSS包含一个执行AP功能的站点，那么可以使用由PCF算法执行的集中访问方法。这种方法保证业务服务的优先级。这样，AP执行**点协调器**（point coordinator）（PC）[⊖]的功能（即介质仲裁器）。

802.11网络的PCF与DCF同时存在。它们使用3种类型IFS进行协调（图14-13）。

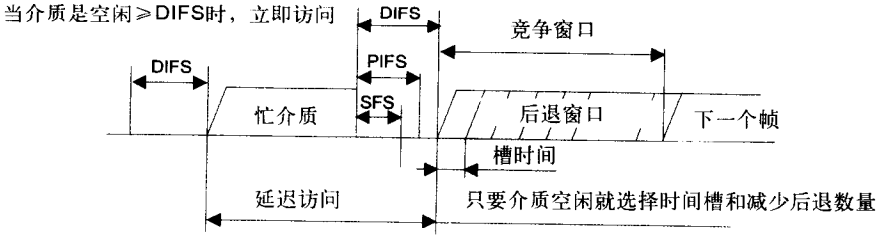


图14-13 PCF和DCF共存

介质释放后，每个站点计算介质的空闲时间，并与三种时间间隔值相比较：

- 短IFS（SIFS）
- 点协调功能 IFS（PIFS）
- 分布式协调功能 IFS（DIFS）

仅当介质空闲时间大于或等于DIFS时，介质捕获过程才使用前面描述的DCF算法。这意味着当描述DCF算法操作时，DIFS（即三种可能的时间间隔的最大者）必须被用作IFS。这给了DCF方法最低优先级。

最小值SIFS的目的是为CTS或ACK帧获得最高优先级的介质捕获，当开始传输帧后继续或完成传输。

PIFS大于SIFS但却小于DIFS。介质仲裁器（即，PC）使用的时间间隔等于DIFS和PIFS的差。在这个间隔期间，它发送一个特殊的信标帧，通知所有站点**非竞争时段**（contention-free period）开始。收到信标帧后，想要使用DCF算法捕获介质的站点不能再这样做。为了捕获介质，它必须等待非竞争时段结束。这个周期的持续时间在信标帧中声明，尽管如果站点没有延迟敏感业务时，它可能提前结束。这样，PC发送CF-END帧，此后DCF访问方法开始执行，假定DIFS已经结束。

在非竞争时段内，PC使用轮询程序给每个加入PCF的站点传输帧的可能。为实现它，PC交替发送一个特殊的CF-POLL帧给每个加入的站点，这样就给了它使用介质的机会。收到这个帧后，站点用CF-ACK+DATA帧应答，确定接收到CF-POLL帧并同时发送数据（通过传输到PC地址或者直接到目的地址）。

为了确保异步业务总能获得部分带宽，非竞争时段的持续时间是有限制的。当某个时间周期结束时，PC发送CF-End帧，竞争时段开始。

任何站都能加入PCF。为此，当连接到介质时（即执行连接程序），必须支持这种业务。

14.4.5 安全

IEEE 802.11的开发者设定目标确保使用无线LAN传输数据的安全性等同于使用有线LAN传输数据的安全性，例如以太网。

⊖ 这部分，PC代表点协调器。

有线以太网的描述中没有特别的方法确保数据安全。以太网标准不执行用户鉴别或数据加密。然而，相比无线LAN，有线以太网更好地防止了非授权访问或机密破解。就因为它们是有线的：入侵者必须物理连接到访问的有线网络。为此，入侵者必须渗透到有孔的建筑物并且连接要攻击的计算机。这个动作可以被观察到并被制止，尽管它能获得非授权访问有线LAN。

无线LAN中，更容易执行非授权访问。在这样的LAN范围内就行。为了成功渗入LAN，必须进入这种LAN运行的建筑物。也需要物理连接介质，因为访问者会毫不怀疑地接收数据。只要手提包中有一个时髦的笔记本就足够了。

802.11提供了安全措施可以提升无线LAN的安全性，使其达到一般有线LAN的水平。因此，802.11网络的主要数据安全协议的名称是**有线对等加密**（wired equivalent privacy，WEP）。它允许使用无线介质对发送的数据加密，这就确保了机密性。无线网络的另一种安全算法是鉴别算法——通过确认唯一性只允许授权的用户登录上网。但是，802.11的安全措施经常受到批评，因为它不能提供像其他标准的安全措施那样的可靠数据传输。例如，发现加密的802.11流量后，资深的入侵者会在24小时内将信息解密。因此，802.11i工作组正在为802.11网络的数据保护开发更强大的标准。

14.5　PAN与蓝牙

14.5.1　PAN的特殊性质

个人区域网（Personal Area Network，PAN）的目的是单个用户的设备在短距离范围内的通信，通常为10m，如笔记本、移动电话、打印机、个人数字助理（PDA）、电视机和许多高技术家庭的装置，如冰箱。

PAN必须能够提供固定访问（即，房间内部）和移动访问（即，当户主携带这些设备在房间、大楼和城市间移动时）。

PAN在许多方面与LAN类似，但它们也有一些特殊的特点：

- 想要加入PAN的许多设备比计算机这个典型的节点简单的多（*mush simpler*），另外，这些设备通常体积小而廉价。因此，PAN标准必须考虑PAN的运行应该是低能耗的廉价技术方案。
- PAN的覆盖范围小于LAN。对于PAN节点间的相互作用，几米的距离足够了。
- 安全的严格要求（Stringent requirements to security）。用户携带的个人设备必须能在各种环境下工作。有时，它们要和其他LAN设备通信。例如，用户在某地会见同事或朋友并且他们决定交换彼此存储在其PDA内地址簿中的地址。另一种情况，这样的交互是不希望有的，因为会导致机密信息泄漏。因此，PAN协议必须确保有在移动环境下各种设备的鉴别和数据加密的方法。
- 小设备互联时，相比将打印机连接到计算机或集中器，取消电缆的需要更为明显。所以，*PAN比LAN更加趋于使用无线方案*。
- 如果用户经常携带PAN设备，它不应损害用户的健康。因此，这种设备发射的信号必须是低功耗的（*low powered*），适合不超过100毫瓦。（一般的蜂窝电话发射信号的范围是600毫瓦到3瓦）。

现在，最流行的PAN技术是蓝牙，它能保证八个设备的互操作，使用2.4MHz的共享介质速率可达723Kb/s。

14.5.2　蓝牙的体系结构

爱立信公司组织的**蓝牙特别兴趣小组**（Bluetooth Special Interest Group，Bluetooth SIG）开发了蓝牙标准。IEEE 802.5工作组采纳了蓝牙标准使其成为IEEE 802标准的一般架构。

蓝牙技术使用**微微网**（piconet）的概念。概念的名字说明这种网络的覆盖区域很小，蓝牙设

备发送端的功率范围可达10m到100m.

一个微微网可以连接255个设备。但是，任何时候只有8个设备是活动的并且可以执行数据交换。微微网中的一个设备是主设备（master），其他设备都是从设备（slaves）（图14-14）。

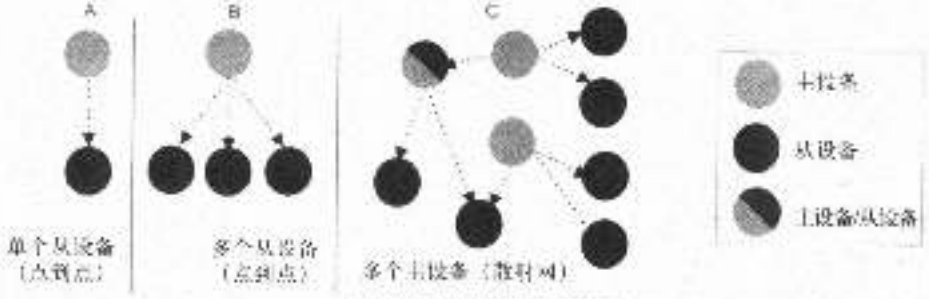


图14-14 微微网和散射网

主设备负责提供访问微微网的共享介质，出现在开放的2.4GHz范围。这种体系允许从设备（如无线电耳机）使用简单协议并由计算机承担复杂的网络管理功能，它也很可能成为网络的主设备。

这样，每个微微网都有1个主设备和7个活动的从设备。1个活动的从设备仅能和主设备交换数据。从设备与设备之间不能直接交换数据。除7个活动的从设备外，当前微微网的所有从设备必须执行低能耗的PARK模式，这种模式下，从设备周期性地监听主设备的命令以便转为活动状态。

主设备负责访问**微微网共享介质**（piconet shared medium），存在于开放的2.4GHz的频率范围内。共享介质以1Mb/s的速率发送数据。但是由于分组头部和跳频的消耗，介质的有效信息速率不超过777Kb/s。主设备用时分复用（TDM）技术将带宽分给7个从设备。

这样的体系可以让从设备使用简单的协议（如耳机），由计算机承担复杂的网络管理功能。它也很可能成为这个微微网的主设备。

连接微微网的过程是动态的。微微网的主设备通过轮询周期性地收集设备上的信息，这些设备处在微微网的区域内。此过程就是询问。检测到新设备后，主设备就与设备执行协商程序。如果从设备的目的是连接主设备所在的微微网（这意味着设备执行了鉴别程序并属于它的可许设备之列），那么主设备将新设备连接到它的网络。

说明 蓝牙网络的安全性是通过设备鉴别和流量加密确保。蓝牙协议确保有比802.11标准的WEP协议更高的保护级别。

不同的区域内执行数据交换的几个微微网形成一个**散射网**（scatternet）。形成散射网的微微网可以相互作用，因为同个节点可以同时成为几个微微网的一部分。这样的设备称为网桥（*bridge*）。散射网与标准802.11 ESS的不同，无散射网中设备能分离为网络（精确说是BSS）和生成AP。散射网中，每个节点在一个微微网中是主设备，在另一个微微网中可能是从设备。

为了防止不同微微网间的信号干扰，每个主设备使用自己的跳频序列。不同的跳频序列使微微网的互操作复杂化。为了完成互操作，和当网桥的设备必须成为每个微微网的一部分并改变它的频率序列。

尽管冲突不太可能发生，但不同微微网的设备选择同样的频率信道操作时，冲突就会产生。它发生的概率很小，因为小区域的微微网数量很少。

按照这个标准的描述，散射网在FHSS的基础上执行CDMA。

为确保可靠的数据传输，蓝牙使用FEC。发送数据时，使用确认信号确定帧的接收。FEC编码不是强制的方法。

蓝牙网络使用不同的方法传送下列两种信息：

- 延时敏感业务，网络支持**面向连接同步链路**（synchronous connection-orientioned link，SCO）。对SCO信道，为连接的所有时间预留带宽。SCO信道通常以64Kb/s的速率传输语音业务。
- 支持弹性业务的**异步无连接链路**（asynchronous connectionless link，ACL）。对于ACL信道，带宽按照从设备的要求和主设备的需要分配。ACL信道目的是为不同速率的计算机业务服务。

14.5.3　蓝牙协议栈

蓝牙是为个人电子设备的单独使用而原创的多功能完整技术。为此，它支持一个完整的协议栈，包括自己的应用协议。这是和以前讨论的技术的主要不同，例如以太网或IEEE 802.11，它们只执行物理层和数据链路层功能。

蓝牙开发者的目的是执行这种技术的各种简单设备不能（实际上也不需要）支持TCP/IP栈，这能解释蓝牙内部固定的应用协议。

人们在尝试开发移动电话和耳机间互作用的标准时，导致了蓝牙的出现。很明显，使用复杂的协议如FTP或HTTP完成这个任务是没有意义的。

结果，开发出了精妙的协议组，除此之外还有相当数量的框架出现。

框架（Profile）执行特殊任务的特殊协议。例如，有的框架是关于计算机或移动电话和无线耳机（耳机数据）间的相互作用。还有一个文件传输框架机一个模拟RS-232端口的框架。文件传输框架是传输文件的设备（尽管未来难以预测，但耳机可能不会用它）使用的。

当蓝牙标准和IEEE 802标准体系一致时，IEEE 802.15.1工作组仅限于研究蓝牙核心协议（*Bluetooth core protocol*），它对应于物理层和MAC层功能（图14-15）。

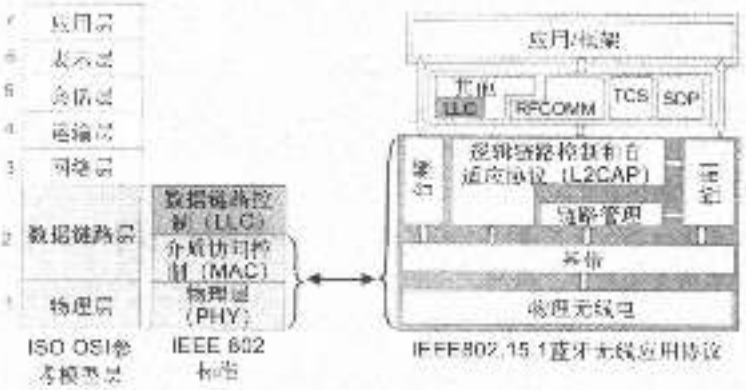


图14-15　蓝牙协议、OSI模型和IEEE802标准的对应

- 物理无线电层（*Physical Radio layer*）描述了用于信息传输的信号频率和功率。
- 基带层（*Baseband layer*）负责组织无线电介质的链路。这个层的责任包括选择跳频序列，同步微微网设备并通过建立的SCO和ACL链路形成和发送帧。蓝牙帧的长度是可变的：数据字段可以包含0到2 774位（343字节），对于语音传输使用固定长度的帧，其数据字段为240位（30字节）。
- 链路管理（*Link Manager*）负责设备鉴别和业务加密。除此之外，它还控制设备状态，例如从设备转变为主设备。
- 逻辑链路控制和自适应层（*Logical Link Control Adapation Layer*，*L2CAP*）是蓝牙核心协议的高一层。当设备发送数据时才用这个协议。语音业务绕过这个协议直接到达基带层。

L2CAP层接收来自高层的64KB数据段并将它们分割成小的帧送到基带层，收到帧后，L2CAP层把这些帧组装成原始段并发送到高层协议。

- 语音层（*Audio layer*）通过SCO信道传输语音。该层使用脉冲编码调制（PCM）的编码，定义了语音信道的速率是64Kb/s。
- 控制层（*Control layer*）传送的所有信息是关于连接外部单元的状态和接收及发数据状态的命令和外部单元的配置。

14.5.4 蓝牙帧

共享介质是2.4GHz频段的一组FHSS频率信道序列。每个频段的带宽是1MHz。信道的数量是79（美国、大多数欧洲国家和大部分其他国家）或23（西班牙、法国和日本）。

码片率是1 600Hz，码片周期是625μs。主设备利用TDM技术划分共享介质，将系统中每个频率信道花费的时间（即625μs）作为一个时间槽。信息在1MHz的时钟频率下用BFSK调制编码，结果，位速率是1Mb/s。

单个时间槽期间，蓝牙理论可传送625位，但是，并非所有的位都用于传送用户信息。当跳到另一个频率时，网络设备需要一些时间进行同步，因此只有625位中的366位用于传送信息帧。

信息帧可以占用1、3、或5个时间槽。当帧占用超过1个时间槽，在帧传输的整个时间内信道频率保持不变。这样，同步的额外开销会减小。包含五个连续时间槽的帧大小是2 870位（数据字段占2 744位）。

说明 只有数据帧（即，ACL信道帧）能包含多个时间槽；传送语音数据的帧（即，SCO信道帧）仅包含一个时间槽。

考虑包含单个时间槽的帧格式——366位（图14-16）。

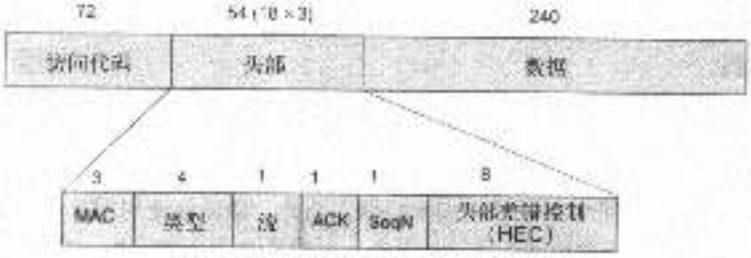


图14-16 包含单个时间槽的蓝牙帧格式

组成帧的366位：

- 240位分配给数据字段（*data field*）。
- 72位被访问代码（*access code*）占用。访问代码用于鉴别微微网。每个蓝牙设备有一个全球唯一的6字节地址。为鉴别微微网，使用主机带一地址的二个最低有效字节。形成哨时，每个设备把这二个字节放到访问代码字段。并用1/3 FEC位补足它们（缩写1/3指1位的信息转变成3位的代码）。如果主设备或从设备收到包含无效访问代码的帧，就丢弃它们，因为这些帧可能已被其他微微网收到。
- 54位用于帧头部（*frame header*），帧头部包含MAC地址、单个位的接收确认标记、帧类型和其他标记。MAC地址是3位，它是7个从设备之一的临时地址，000作为广播地址。头部信息也使用1/3 FEC码传送。

包含3个和5个时间槽的帧格式的不同之处仅仅是数据字段大小。数据字段的信息编码采用1/3FEC或2/3FEC，或者不使用FEC发送。

14.5.5　蓝牙如何运作

考虑一个微微网运作的例子。假定这个微微网包含1个主设备和3个活动的从设备。为了简单，假定所有设备使用的就只与一个时间槽。图14-17显示了主设备如何在微微网的成员中分配从设备。

为确保信息交换的全双工模式，主设备总是分配给每个信道两个从设备。第一个时间槽用于一个主设备向另一个从设备传输数据，第二个时间槽用于反向传输。

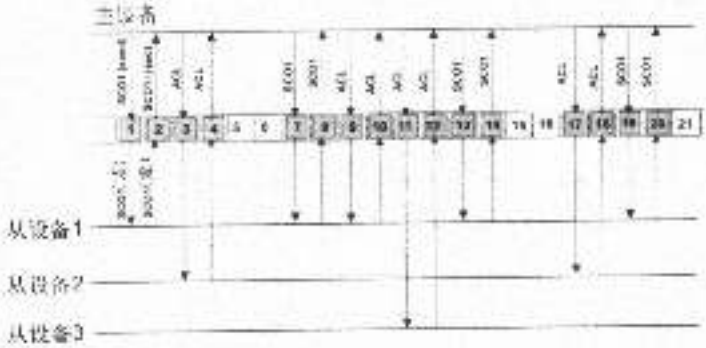


图14-17　共享介质

在如图14-17所示的例子中，主设备和从设备1间有一个SCO信道。按照语音信息编码使用的FEC方法分配给SCO信道固定的带宽。

1）如果不使用FEC，如图14-17所示，每个第三对时间槽分配给SCO信道。这种时间槽分布确保每个方向的传输流量是64Kb/s。你可以检查它，PCM编码器对语音数据以8KHz的频率采样（125μs的周期），同一个方向传送的SCO信道帧每6个时间槽重复一次，帧间的周期是6×625=3 750μs，相应地，信道的信息速率（一个方向）是240/(3 750×10⁻⁶)=64Kb/s。

2）使用2/3编码，被数据字段包含20个采样而不是30个，确保速率是64Kb/s，每个第二对时间槽分配给SCO信道。

3）最后，1/3 FEC编码使每一个帧传送10个语音采样，占用共享介质的所有时间槽。

计算显示，微微网（可能有不同的从设备）内只能存在不超过3个SCO。尽管如此，当信道不使用FEC编码或少位错误时才有这种可能性。使用FEC将SCO的信道数量减少到两个甚至一个。

形成SCO信道后剩余的带宽用于传送同步数据。为此，微微网使用ACL信道。这是个点到多点的信道，连接主设备到所有微微网的活动从设备。不需要建立这种信道，因为它总是存在的。

主设备周期性地轮询从设备，检测它们是否需要发送同步数据。为此，使用含有特定设备的MAC地址的特殊POLL帧。如果主设备有这个设备的数据，它能把数据传输和轮询集中到单个帧中。

图14-17显示了主设备利用时间槽3和4与从设备2交换帧，时间槽9和10用于与从设备1交换帧，时间槽11和12用于从设备3。访问ACL信道时轮询的方法消除了冲突。但是，访问这个信道的速率不是固定的；相反，依赖于需要发送异步数据的设备数量。

总体，蓝牙网络包含电路交换（SCO信道）和分组交换（ACL信道）。

如果SCO信道不用于蓝牙网络，整个带宽分配给ACL信道，当使用5个时间槽组成的帧时，每个方向的最大数据传输速率是432.6Kb/s（不使用FEC）。ACL信道带宽的异步划分也是可能的。这样，一个方向的最大速率可达723.2Kb/s，反方向的速率是57.6Kb/s。这些是ACL信道的速率，不是来自特定设备的流数据率。当多个设备共享ACL信道时，这个速率在所有涉及到的设备间划分。

14.6　共享介质LAN的设备

具有网络适配器的集中器和电缆系统是构建共享介质LAN最低要求的设备。显然，这样的网

络不能太大，因为假如网络节点的数量增长很快，那么共享介质就会变成瓶颈。因此，集中器和网络适配器可以建成网络的一个小的基本段，它们之间通过交换机、网桥和路由器连接。

14.6.1 网络适配器的主要功能

带有驱动程序的**网络接口卡**（network interface，NIC）执行网络端节点（计算机）中的OSI模型的第二数据链路层。更确切地说，在网络操作系统中，适配器-驱动程序对只执行物理层和MAC层功能；LLC层的功能通常由所有驱动程序和网络适配器共同的OS模块执行。例如，在Windows XP中，LLC层是在对所有NIC驱动器公共的网络驱动程序接口规范（NICF）模块中执行，它是由特殊驱动程序支持的独立技术。

网络适配器和它的驱动程序联合执行两种操作：接收和传送帧。

从计算机到电缆的帧传输（*frame transmission*）包含下列阶段：

- 通过服务接口接收LLC帧，其信息中含有MAC层地址。在计算机内，协议之间使用RAM缓存器互相作用。高层协议使用OSI/O子系统找回通过网络传送的数据，这些数据来自磁盘或文件缓存器，然后把数据载入RAM缓存器。
- 将MAC帧格式化后封装到LLC帧中。它包括填充的源和目的地址及计算校验和。
- 形成代码符号，假定使用冗余码如4B/5B和扰码得到平滑的信号频谱。并非所有的协议都执行这个阶段。例如，10Mb/s以太网就没有。
- 按照使用的线路编码曼彻斯特码，NRZI，MLT-3等等发送信号到电缆。

从电缆接收帧包括下列阶段：

- 从电缆信号接收编码的位流。
- 将信号与噪声分离。这可以由特殊电路或数字信号处理器执行。因此，适配器的接收端得到一个位序列，它与发送端发出的序列一致。
- 如果数据在发送前被扰乱，使用消扰器将它们发出。这个操作后，发送端将发出的代码符号存入适配器。
- 检测帧的校验和。如果校验和不正确，就将帧丢弃，并且将一个特殊的错误代码通过服务接口发送到LLC协议。如果校验和正确，就从MAC帧中重新获得LLC帧并通过服务接口发送到LLC协议。

标准没有定义网络适配器和它的驱动程序的责任划分。因此，每个制造商都自由地解决这个问题。因此，网络适配器分为两类：客户端计算机的适配器和服务器的适配器。

客户端计算机（*client computer*）的适配器中，大部分任务分配给驱动程序。这样，适配器就不复杂，价格低。它的缺点是CPU负载过大，在这种情况下，必须执行把数据从RAM缓存器传输到网络这样的常规操作。

服务器（*server*）的网络适配器装备了嵌入式处理器，处理器将数据从RAM传输到网络以及反向传输涉及的大部分任务都由它执行。

按照适配器执行的协议，分为以太网适配器、令牌环适配器和FDDI适配器等等。因为快速以太网使用自动商量程序根据集中器的能力自动选择网络适配器的操作速度，所以许多以太网适配器支持两种操作速度，它们名字的前缀是10/100。

网络适配器使用管道的方法处理帧。按照这种方法，处理从计算机RAM接收的帧和传输帧到网络是同时发生的。这样，收到帧的几个起始字节后，适配器就开始发送它们。RAM-适配器-物理链路-适配器-RAM串的性能有相当大的提高（25~55%）。这种方法对传输开始的阀值非常敏感。（即，开始实际传输之前，帧字节的数量必须载入适配器）。网络适配器通过分析介质和计算阀值对这个参数实行自我调整，不需要网络管理员的介入。自我调整确保计算机内部总线和它的IRQ和

DMA装置的特殊结合达到最大性能。

网络适配器基于专用集成电路（ASIC），ASIC提高了它们的性能和可靠性同时降低了成本。

说明　提高存储器和适配器间信道的操作速率对于整个网络性能的提高是非常重要的，因为帧的传输路由可能包括集中器、交换机、路由器和WAN链路，帧的传输速率取决于路由中最慢设备的性能。所以，如果服务器或客户端的网络适配器操作很慢，就算最快的网络设备也不能提高整个网络的操作速度。

现在生产的网络适配器可归类为4G（第四代）适配器。它们必须包含执行MAC层功能的专用集成电路（ASIC）芯片，和高层的功能。这些功能集合包括支持远程监视代理，帧优先级划分方法和远程控制功能。服务器的网络适配器几乎都包含功能强大的嵌入式处理器，它能减少CPU的负载。

14.6.2　集中器的主要功能

实际上，当代所有的LAN技术定义了一个有三种名字的设备，这些名字可以互换：集中器、集线器或中继器。按照不同的应用领域，它的设计和功能集有相当大的变化。只有它的主要功能保持不变，也就是，按照相关标准定义的算法，或者在所有端口（如以太网标准定义）或仅在某些端口转发帧（*repeating the frame*）。

通常，集中器有多个端口，网络端节点和计算机使用独立的物理电缆段连接它们。集中器把网络的独立物理段连成公共的共享介质，依照前面介绍的LAN协议访问它。这些协议是：以太网、令牌环等等。因为共享介质的访问逻辑强烈依赖于技术，制造特殊的集中器能适应各种流行的技术。

每个集中器执行一个主要功能（*main function*），由其支持的适当技术标准定义。

除了主要功能外，集中器执行多种附加功能（*add-on function*），它们或者是标准定义的或者是可选择的性能。例如，令牌环集中器可以断开不正确操作的端口并且转换到预留的环，尽管标准并没有描述这种功能的性能。集中器被认为是便捷的设备，它能执行辅助功能从而简化网络维护和控制。

举一个以太网集中器的例子，考虑集中器的主要功能的特别执行的性质。

在以太网中，把多个同轴电缆的物理段连接成单个的共享介质的设备使用了很长时间。基于它们的主要功能——将输入端收到的信号从所有输出端转发出去——这些设备就是我们所知的以太网中继器。在基于同轴电缆的网络中，**双端口中继器**（two-port repeater）是最普遍的，只连接两个电缆段。所以，术语集中器很少用于它们。

随着10Base-T双绞线的使用，中继器成为了以太网网络的必备部分，因为没有它们，通信只能在两个网络节点间进行。基于双绞线的**多端口以太网中继器**（Multiport Ethernet repeater）又称为集中器或集线器，这是因为单个设备集中连接许多网络节点。图14-18显示了一个典型的以太网集中器，目的是创建小的共享介质段。它有16个具备RJ-45连接器的10Base-T端口和连接外部收发器的单个AUI端口，结果，连接这个端口的收发器是使用同轴电缆或光纤。使用这种收发器，集中器连接着几个集中器的主干电缆上，通常收发器使用同轴电缆或光纤连接到这个端口。使用收发器。站点以同样的方式在离集中器100m或更远的地方连接。

说明　为把10Base-T集中器连接成一个层次系统，同样的端口可被用于连接端用户工作站。但是，这样系统中要考虑一种特殊的情况：连接网络适配器的普通RJ-45端口和称为能交错配置的介质独立接口（MDI-X，X代表交错配置）反转连接，这样就能使用标准连接电缆将网络适配器连接到集中器，不需要交叉连接（图14-19）。当集中器使用标准MDI-X端口连接时，必须使用带有交叉连接对线的非标准电缆。因此，一些制造商提供具有专用MDI端口的集中

器，而不需要使用交叉对线。这样两个集中器可以按级联的方式连接，使用一个集中器的MDI-X端口直接通过电缆连接另一个为MDI端口设备无关的位置。集中器的同一端口可用作MDI-X和MDI端口，如图14-19下面部分所示。

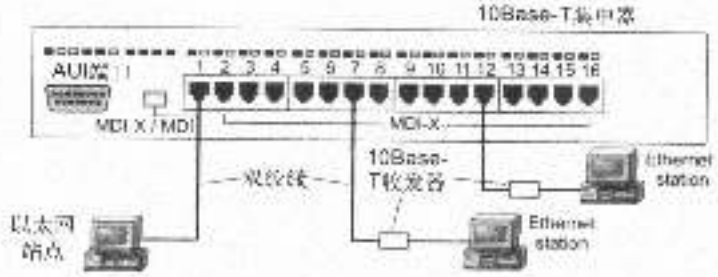


图14-18 以太网集中器

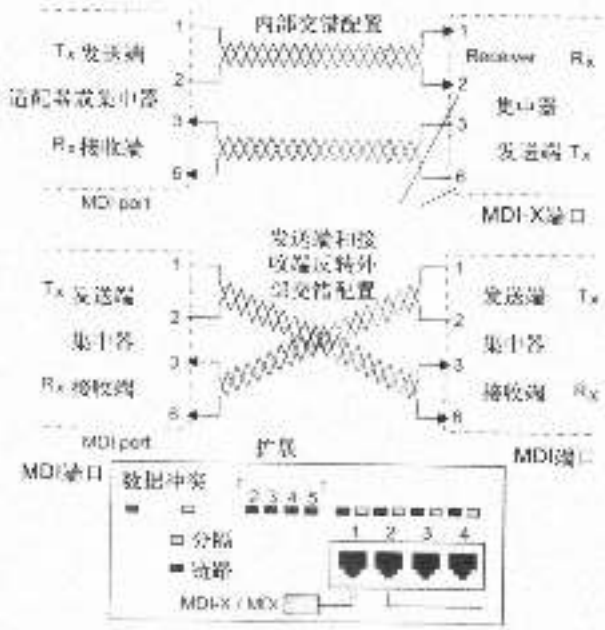


图14-19 基于双绞线的站点-集中器和集中器-站点连接

使用四集线器规则（four hubs rule）时，从不同的情况考虑多端口以太网中继器-集中器，大多数模型中，所有端口都连接单个集中器单元；当信号通过两个端口时，中继器单元只引入一次延时，因此，这样的集中器被认为是按照四集线器规则强加限制的单个集中器。但是，其他中继器模式在其中继器单元有多个端口。这样，每个中继器单元认为是单个集中器，使用四集线器规则时必须单独考虑。

但是，执行主要集中器功能的差异是无关紧要的。相比之下，执行集中器辅助功能的差异值得考虑。

14.6.3 自动分隔

自动分隔（Autopartitioning）是有用的集中器功能，它能使集中器断开与不正确运行的端口的连接，将网络的所有其他部分与不正确运行的节点产生的问题隔开[⊖]。以太网和快速以太网标准

⊖ 在HDDI集中器中，由于它在协议中定义，所以这个功能对大部分错误情况是主要的。

端口断开的主要原因，是不能响应每隔16ms发送到所有端口的链路测试脉冲序列。这样，故障端口被转到断开状态。但是，当设备恢复时链路测试脉冲将继续发送到其端口，这会自动运行。

考虑以太网和快速以太网断开端口的情况：

- *帧级别的错误（errors at the frame level）*。如果通过端口的错误帧的密度超过了预定阈值，那么断口被断开。假定在预定时间间隔内没有发生错误，端口又被重新连接。这些错误可能包含不正确的校验和、无效帧长度（超过1 518字节或少于64字节）或不正确的帧头部。
- *多重冲突（Multiple collision）*。如果集中器记录了同一端口成为冲突源超过60次，那么端口被断开。过些时间，端口又被连上。
- *超长传输（Lengthy transmission，jabber）*。像网络适配器一样，集中器控制单个帧通过端口的时间。如果这个时间超过传输最大长度帧的时间间隔两倍，端口被断开。

14.6.4　反相链路的支持

因为只有FDDI定义了使用集中器中的反相链路，所以集中器的开发者支持集中器的这种功能仅作为可选的技术。例如，以太网集中器能形成没有环的分层次链路。因此，断开端口间的预留链路以防止破坏网络操作逻辑。通常，在配置集中器时，网络管理员必须决定哪个是主要端口，哪个是预留端口（图14-20）。如果端口由于某些原因断开（也就是，自动分隔机制随即启用），那么集中器激活预留端口。

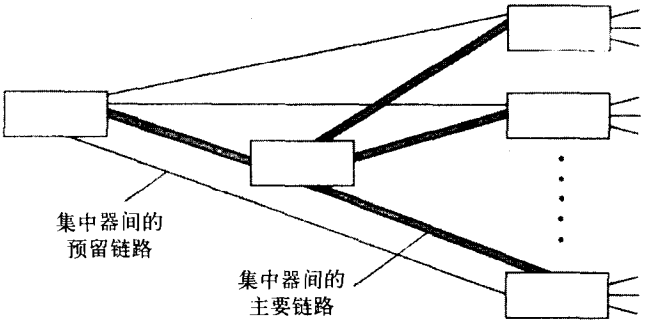


图14-20　以太网集中器间的预留链路

某些集中器模型中，只对最重要的基于光纤电缆链路允许使用端口预留机制。其他模型中任何端口都可能预留。

14.6.5　保护以防未授权访问

共享介质使未授权访问侦听和访问传输的数据非常容易。为实现它，可以将装有协议分析器副本的计算机连到一个空闲集中器的连接器上，将所有通过网络的流量存储到硬盘文件中。此后，有可能得到需要的信息。

集中器的制造商提供了一些在共享介质中保护数据的方法。

最简单的保护方法是分配允许的MAC地址给集中器端口。在标准以太网集中器中，端口没有MAC地址。数据保护就是人工分配特别的MAC地址给每个集中器端口。MAC地址是允许连接端口的站点地址。例如，图14-21中，集中器的第一个端口分配了特别的MAC地址（简单地表示为123）。相同MAC地址的计算机能使用这个端口与网络正常通信。如果入侵者断开计算机并且连接其他计算机，那么新的计算机启动后集中器会发现，从新计算机进入网络的帧的源地址变了（即，变到789）。由于这个地址对第一个端口是无效的，所以这些帧会溢出，端口被断开，记录了一次安全事件。

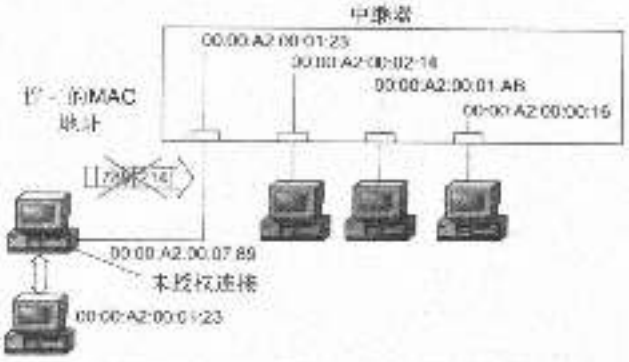


图14-21 端口隔离：只允许从预定了MAC地址的站点传输帧

为了执行集中器数据保护的方法，集中器必须重新配置。为此，集中器必须配置控制部件。这样的集中器通常称为智能集中器。控制部件是内置软件的小型计算部件。为了确保管理员能够与控制部件通信，集中器必须有控制台端口（经常是RS-232端口），它连接一个终端或是装配终端仿真程序的PC。当终端连接控制台端口时，控制部件显示一个对话框，网络管理员可以在其中输入MAC地址。控制部件能支持其他配置操作，如人工连接或断开端口。为此，控制部件在其终端屏幕显示一些菜单形式。使用这些菜单，网络管理员可以选择需要的动作。

另一种防止数据未授权访问的方法是在集中器中加密。但是，真正的加密需要强大的计算机功能。因此，对于没有帧级存储的集中器，传输中的帧加密是个大问题。不是真正的加密，集中器只是对发送到端口的分组中的数据字段随机扰乱。这些端口的地址与分组的目的地址不同。这种方法保持了随机访问介质的逻辑，因为所有站点会发现介质正忙于传输信息帧。但是，只有发送的就想要到达的目的端口的站点可以正确解释数据字段的内容（图14-22）。为了使用这种方法，集中器必须有连接到它端口的所有站点的MAC地址。发送到目的节点以外的其他站点的帧，其数据字段用零填充。

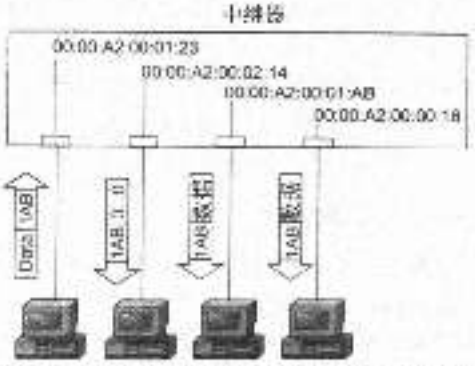


图14-22 到达非接收站点的帧中的数据字段被扰乱

14.6.6 多段集中器

为什么有的集中器配备这么多数量的端口，192或240？在大量站点中把介质分为10或16Mb/s有意义吗？10或15年前，这些问题在某种意义上说是肯定的。例如，网络中的计算机使用介质只是发送小的邮件报文或者拷贝小的文档文件，这是可以实现的。今天，没有多少这样的网络，甚至五台计算机完全能承载一个以太网段。

那么为什么需要一个集中器具备许多端口，甚至无如其受到每个站点的带宽限制，实际上不可能使用全部的端口？

答案是这样的。集中器有多个没有连接的内部总线。这些总线用于创建多个共享介质。例如，图14-23所示的集中器有三个内部以太网总线。如果这种集中器有72个端口，每个端口都能连接三个内部总线中的任意一个。图14-23的结构显示前两个计算机连接以太网3总线，第三和第四站点连接以太网1总线。前两台计算机形成一个共享段，第三和第四站点形成另一个共享段。

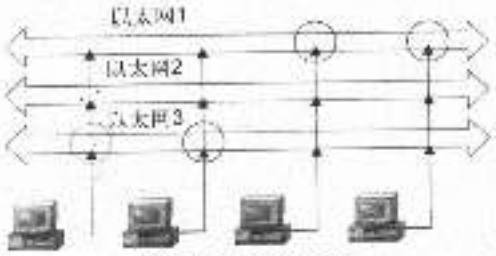


图14-23　多段集中器

连接到不同段的计算机不能使用集中器进行通信，因为它的内部总线不相互连接。

多段集中器用于连接能够容易变化的共享段。大多数多段集中器可以通过程序连接端口到它的内部总线。例如，使用控制台端口名本地配置。类比，网络管理员可以连接用户的计算机到任意集中器端口，并且使用集中器配置程序控制每个段的结构。例如，如果某段1拥塞，连接它的计算机可以分配在集中器的保留段中。

编程改变端口和集中器内部总线的连接称为**配置交换**（configuration switching）。

说明　配置交换与网桥和交换机执行的分组交换是不同的。

多段集中器是大规模网络的设计基础。段的互联需要使用其他设备：网桥、交换机或路由器。这些网络互联设备必须连接到有不同内部总线的多段集中器的多个端口。这个设备的主要目的是当帧在不同段间转发时，就好像使用独立的集中器。

对于大型网络，多段集中器起着智能箱柜的角色。它按程序创建新的连接，通过改变内部设备的配置而不是机械地连接电缆插头到另一个端口。

14.6.7　集中器设计

集中器的应用领域对其设计有着重要的影响。通常，工作组集中器被认为是一种有固定端口数量的设备，协作集中器是基于底板的模块化设备，部门级集中器有一个栈结构。这样的划分并不严格。模块化集中器也能用作企业级设备。

有固定端口数量的集中器（Concentrators with a fixed number of port）设计最简单。这种设备是有必备元件（端口、指示器、控制器和电源部分）的独立部件。这些元件不能被代替。通常，这样的一个集中器的所有端口支持一种传输介质；端口总数量是从4到48。一个端口专用于连接集中器到网络主干或者连接集中器。（通常，AUI端口用于这个目的；使用适当的收发器可以使集中器实际连接任何物理传输介质）。

模块化集中器（Modular concentrator）是通过在公共底盘上安装固定数量的端口的独立模块来实现。底盘上有内部总线用于把独立模块连接成普通的中继器。通常，这样的中继器是多段的，其中单个模块集中器内有多个相互连的中继器。一个模块集中器可能存在几种类型的模块，其不同之处是端口数量和支持的物理介质。模块化集中器可以选择更精确的集中器配置。它们也有灵活性，应对网络配置变化时成本低。

因为企业模块化集中器执行重要任务，所以配备了控制单元、热量控制系统、冗余电源部件和可以自动替代的模块。

当业务需求在网络配置的初始阶段仅安装了一个或两个模块时，基于底盘的集中器的最大缺点是价格昂贵。底盘价格高是由于配有其他的所有设备，例如，冗余电源部件。因此，**栈集中器**（stack concentrator）对于中等大小的网络是最流行的。

与有固定数量端口的集中器类似，栈集中器配备独立的部件，不可能替换单个模块，多种以

太网堆集中器的典型例子，如图14-24所示。

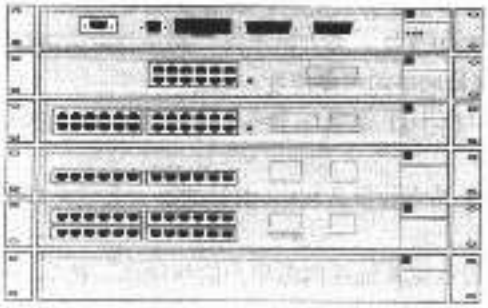


图14-24 以太网栈集中器

栈集中器用特别的端口和电缆连接多个这样的部件使之成为单个中继器（图14-25），它有公共的中继器单元，确保所有信号同步，因此按照以集线器规则认为是单个中继器。如果栈集中器有多条内部总线，将这些总线连接在一起，当这些集中器连接到栈时，对所有栈设备而言这些总线成为共公总线。

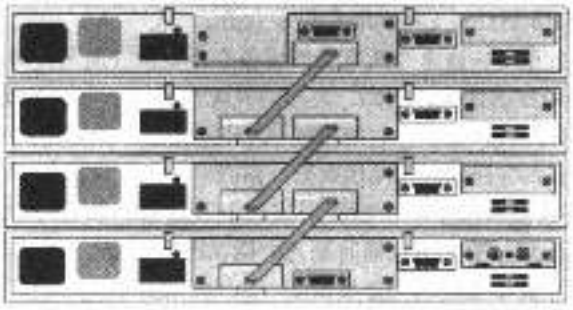


图14-25 后面板上使用特别连接器将栈集中器连成一个单个设备

连接到栈的设备数量可以非常大（通常可达八个，有时更多）。栈集中器支持不同的物理传输介质，使得它们几乎和模块化集中器一样灵活。但是，栈集中器的每个端口的成本较低，因为一个业务可以从单个设备开始而不需要冗余的底盘并且按需添加类似的设备到栈。

同一制造商的栈集中器通常有相同的设计，使其很容易安装在另一个上面，这样就形成单个桌面设备，或者把它们放在公共点上，当组装栈时，可以节省栈中所有设备的公共控制部件，它能作为辅助模块插入共共底盘。另外可以使用公共冗余电源部件提供应急电源。

模块化栈集中器（Modular stack concentrator）是使用特殊的链路将模块化集中器连接到栈中。通常，这样的集中器用于数量少的模块（1~3）。这些集中器集合了两种类型集中器的优点。

这种集中器设计的分类不仅可以用于集中器还可用于其他类型的通信设备：LAN网桥和路由器和WAN网桥和路由器。并非所有设备都像集中器那样与栈部件有紧密的联系。通常，栈设备仅连接公共电源部件和控制单元，主要功能由每个站点设备自动执行。

小结

- 令牌环网使用确定性访问执行令牌传递。这种方法保证每个站点在令牌传递期间访问共享介质。令牌环网的逻辑拓扑是环，物理拓扑是星形。
- 令牌环网执行两种速率（4或16Mb/s），可以使用屏蔽或非屏蔽双绞线也包括光纤作为传输介质。环中最大站点的数量是260。环的最大长度是4Km。环拓扑的应用可以使令牌环网络

确保基本的容错特性。

- 从令牌环到光纤的继承是非常有意义的：两者都使用相同的网络拓扑，都使用令牌传送作为访问介质的方法。FDDI支持高级容错手段，隔开故障的电缆系统或环的工作站点，通过叠加双环到单个环中保持网络的可操作性。
- 光纤分布式数据接口（FDDI）是第一个在LAN中使用光纤并且确保数据传输率为100Mb/s的技术。
- FDDI环的双连接的最大站点数量是500；双环的最大直径是100Km。这使得FDDI不仅可用于LAN还可用于MAN。
- 无线LAN去除了庞大的电缆系统并确保用户的移动性。但是，需要网络架构师解决一系列复杂问题，如有关无线介质高电平噪声特性和未确定的网络覆盖区域。
- IEEE 802.11标准是无线LAN最有前途的标准。802.11物理层有多种不同的规范，差别在频段（2.4或5GHz）和编码方法（FHSS、DSSS或OFDM）。802.11b物理层确保数据传输率达到11Mb/s。
- 802.11的访问方法是结合了冲突避免的随机访问和基于轮询的集中式确定性访问的方法。第一种方法通过分布式协调功能（DCF）算法执行，第二种使用点协调功能（PCF）算法执行。
- 灵活使用DCF和PCF允许QoS支持同步和异步业务。
- 个人区域网（PAN）用于构建单个所有者的设备在小范围（通常，10~100m）内的相互作用，PAN必须确保固定和移动访问，例如，大楼内或者在房间、大楼或城市间移动。
- 当今，蓝牙是最流行的PAN技术。它使用微微网的概念。一个微微网可以包含255个设备，但仅有八个是活动的并可在任何时间交换数据。微微网中一个设备是主设备，其他是从设备。
- 处在相同区域并交换数据的多个微微网形成一个散射网。散射网中的微微网可以通过单个节点（网桥）相互作用，单个节点同时是几个微微网的部分。
- 对于延时敏感业务，蓝牙支持面向连接同步（SCO）链路；对于弹性业务，使用异步无连接（ACL）链路。SCO链路通常用于传送64Kb/s的话音业务，ACL信道用于速率可变的计算机业务，最大速率可达723Kb/s。
- 除了LAN集中器的主要协议功能（将到达所有端口或下一个端口的帧逐位复制）以外，它们还执行多种有用的辅助功能：
  - 自动分隔是辅助功能中最重要的一个，如果集中器检测到连接它的电缆或端节点有问题，就使用自动分隔功能断开这些端口。
  - 通过不允许未知MAC地址连接到集中器端口防止未授权访问从而保护网络。

复习题

1. 描述令牌环使用的介质访问算法。
2. 有源监视器执行什么功能？
3. 如果形成环的某个计算机停止工作，那么令牌环为何还能保持连接？
4. 说明下列技术允许的数据字段的长度：
   - 以太网
   - 令牌环
   - FDDI
   - 蓝牙
5. 令牌环网选择最大令牌周转时间的依据是什么？
6. 令牌环网的哪种元件恢复位流的同步？

7. 早期令牌释放算法的优势是什么？
8. FDDI和令牌环有什么共同的性质，不同的又是什么？
9. FDDI网络的什么元件确保容错？
10. FDDI容错，这表示任何单个电缆损坏的情况下，网络能继续正常运行吗？
11. FDDI环中，复制的电缆损坏的结果是什么？
12. FDDI网中，单连接站（SAS）电缆被破坏会发生什么？
13. IEEE 802.11网络使用什么信号编码方法？
14. DS使用何种类型的介质在BSS间传输数据？
15. 隐藏终端效应的影响是什么？
16. 802.11网络的MAC层如何检测冲突？
17. 802.11网络的站点可以使用AP传输一个帧到属于同样BSS的另外的站点吗？
18. DFC中，把时间周期分割成时间槽来传输帧的目的是什么？选择时间槽持续时间要考虑什么？
19. 为什么PCF的优先级高于DCF？
20. 蓝牙的微微网如何连接成散射网？
21. 为什么蓝牙时间槽不完全使用625位传输帧？
22. 什么情况下，蓝牙帧携带1、2、3个SCO信道的数据？
23. 蓝牙使用什么交换方法？
24. 为什么蓝牙使用主设备——从设备结构？
25. 网络适配器带宽和集中器端口带宽如何影响网络性能？
26. 集中器是怎样支持反相链路？
27. 按照集中器的主要功能——转发信号——集中器被归类为在OSI模型的物理层运行的设备。提供例子说明集中器的辅助功能需要来自高层协议的信息。
28. 模块化集中器和栈集中器有什么不同？
29. 用于集中器互连的特殊端口是什么？

练习题

1. 令牌环网有160个站点并且运行速率是16Mb/s，请估计访问介质的最大等待时间。
2. 一个令牌环网包含100个站点。环的总长度是2 000m。传输率是16Mb/s。令牌保持时间是10ms。每个站点传输帧的固定大小是4 000字节（包含头部）并且整个令牌保持时间传送它的所有帧。计算在这个网中使用早期令牌释放算法产生的增加量。
3. IEEE 802.11和蓝牙网络在同一范围运行。802.11网络使用FHSS物理层规范以1Mb/s的速率传输数据。蓝牙网络执行1 600Hz的标准的片速率，并且802.11网络支持50Hz的码片速率。两个网络都在2.4GHz频段使用79个信道。

   由于两个网络中都使用同样的频率信道，确定每个网络中帧损坏的比例。更确切些地说，考虑蓝牙网络的所有数据在一个时间槽帧的传输，802.11网络使用最大长度的帧。

第15章 交换LAN基础

15.1 引言

共享介质一出现就被用于LAN中。用这种方法来使用共享链路有很多优点，其中之一就是简化了LAN的通信设备。然而，共享介质的使用也不是没有缺点的。共享介质LAN最明显的缺点就是低可延拓性，因为随着LAN节点的增加，将导致分配给每一个节点的带宽成比例地减少。

解决LAN的可延拓性问题的正常方法就是将它分割成一些段，每段表示一个独立的共享介质。这种逻辑分割是由网桥或者交换机实现的。在第3章，我们考虑了逻辑网络结构的原理。这一章，我们将就网桥和交换机的实现算法进行更详细的讨论。

交换LAN（switched LAN）就是将LAN分成若干个逻辑段。只包含一台直接与交换机端口相连的电脑的网络段叫做**微网段**（microsegment）。本质上，微网段已经不是一个共享介质了。相反，它是计算机或交换机端口的发送设备在需要时使用的双向信道，这个信道是不用和其他发送设备分享的。

虽然交换LAN方法要比共享介质LAN更加昂贵，但是它有包括可延拓性在内的许多优点。本章将会提到交换LAN的主要优点。

15.2 使用网桥和交换机的逻辑网络结构

15.2.1 共享介质LAN的优点与不足

当建立一些包含10到30个节点的小型网络的时候，使用基于共享介质的标准技术是一个即经济又有效的方法。这是因为它有以下网络特性：

- **简单的网络拓扑**（Simple network topology）：简单的网络拓扑使得网络节点数目可以非常容易地在合理范围内增加。
- **帧丢失的消除**（Elimination of frame loss）：在通信设备上，缓冲区的溢出会导致帧的丢失。而通过使用共享介质技术，只有前一个帧被接收到后，下一个新的帧才会传送到网络上。介质共享自己的逻辑是强迫产生帧过于频繁的工作站推迟它们的帧发送来达到规范帧流量的目的。这些工作站必须等待直到它们被允许进入介质。这样，流量控制的过程就可以自动执行了。
- **协议的简单性**（Simplicity of protocol）：这点确保了网络适配器、中继器和集线器乃至整个网络的低代价。

然而，连接成百上千个网络节点的大型网络不能只建立在一个共享介质之上的论断仍然是正确的。即便是对于像千兆以太网这样的高速技术这个论断也是正确的。实际上，所有的技术都限制了共享介质中网络的最大距离以及最大节点数。例如，对于以太网家族的所有技术，节点都被限制在1 024个以内；对于令牌环技术，这个数目是260个节点；对于FDDI网络，是500个节点。然而，这不是受到限制的唯一原因。

整个网络不能仅仅基于一个共享介质的根本原因是带宽。

可以使用队列模型对共享介质LAN的过程进行定量描述。其中之一叫做M/M/1模型，这个模型我们在第7章中已讨论过。在这个模型中，共享介质对应于服务器，每个联网的计算机产生的帧

对应于服务请求。服务请求的队列分布在所有联网的计算机中，帧在队列中等待使用介质。

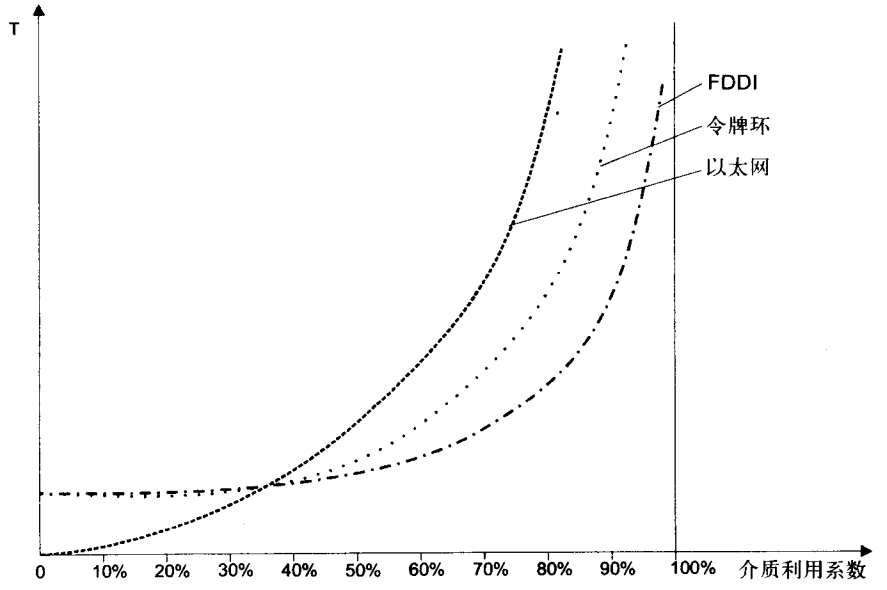


图15-1 以太网、令牌环和FDDI技术的介质访问延迟

M/M/1模型不能充分反映许多共享介质LAN的某些特性，例如以太网冲突。然而，它可以定性地描述介质访问延迟和介质利用系数之间的关系模式。

图15-1展示了使用仿真技术获得的以太网、令牌环和FDDI网络的关系曲线。

从这个图可以看出，所有的技术都有一个相对于负载呈指数级增长的定性的相似的模式。在所有情况下，伴随着共享介质使用的增加，介质访问延迟都有着指数级的增长。然而，它们的阈值有所不同，在阈值处网络行为会发生急剧变化，这里的急剧变化是指一下由线性关系变为指数级的急剧增长。对于所有以太网家族中的技术，这个值是30%~50%（因为冲突的影响），对于令牌环，这个值是60%，对于FDDI，这个值是70%~80%。

由于运行于网络节点上的应用程序类型的不同，共享介质中可以存在的节点数也不同。例如，早些时候，可以认为在以太网环境下，一个共享段中有30个节点是可以接受的。如今，如果网络节点运行多媒体程序或者交换大型数据文件，则一个共享段中只能有5到10个节点。

15.2.2 逻辑网络结构的优点

只使用一个共享介质而产生的限制是可能被克服的，这主要是通过将网络分割成多个共享介质，然后使用特殊的通信设备将这些独立的网络段连接起来，这些特殊的通信设备如网桥、交换机或路由器（图15-2）。

这些设备基于对帧中包含的目标地址的分析在端口之间进行帧传输。网桥和交换机则是基于平面型数据链路层的地址（MAC地址）执行这些帧的传输操作，路由器使用层次型网络层的地址实现这个目的。我们将会在第四部分进一步探讨路由器的操作。这里我们集中探讨网桥和交换机。

我们已经在第3章简要地探讨了逻辑网络结构。在这一节，我们将更加详细地讨论这个问题。在逻辑网络结构有很多作用，其中最主要的是改进了网络性能、灵活性、安全性以及可管理性。

性能改进（performance improvement）。作为一个这种效果的图例，同时也是逻辑结构的主要目的，请看图15-3。这个图展示了一个网桥所连接的两个以太网网段。在这些网段中有一些中继器。在网络被分段前，这个网络的所有节点产生的所有流量共享同一个网络介质。例如，如果这

半连接的不是互联网设备（图中是网桥），而是一个中继器。在考虑网络利用系数时也要考虑到这个网络。假设将从节点i到节点j的平均流量强度记为C_{ij}，那么，分段前网络中需要传输的总流量就是$C_\Sigma = \Sigma C_{ij}$（考虑到这总和是针对所有节点的）。



图15-2　逻辑网络结构

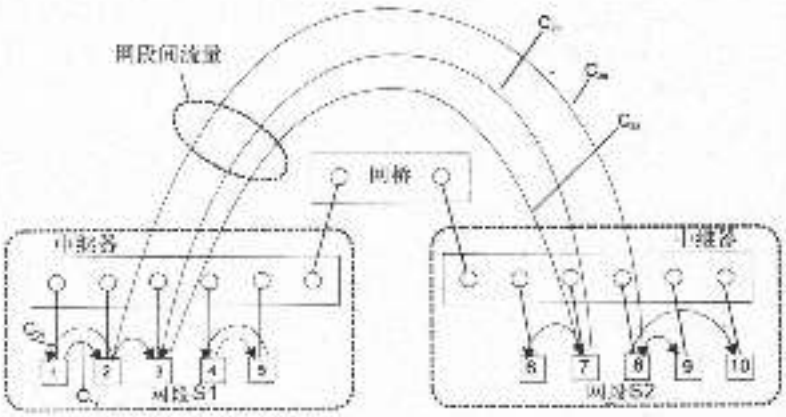


图15-3　分段后网络负载的变化

分段之后，必须既考虑到这个网段的内部流量——即，在一个网段内的各节点间传播的帧，又要考虑网段间的流量，即这个网段的节点发往另一网段节点或另一网段节点发往这个网段节点的帧。

因此，假设把该网段称为S1，则它的网段负载等于$C_{S1}+C_{S1-S2}$，这里C_{S1}是S1网段内部流量，C_{S1-S2}是网段间的流量。为了明确S1网段上的负载是如何变得比最初的网络负载要低，注意到在分段前的总网络负载可以用这个式子表示：$C_\Sigma = C_{S1}+C_{S1\_S2}+C_{S2}$。所以，S1在分段后的负载就等于$C_\Sigma - C_{S2}$，也就是说它减去了S2网段内部的流量。网段S2也是同样道理。因此，通过图15-1我们可以发现，网段的延时减少了，每个节点的有效带宽增加了。

前面我们曾提到，网络分段几乎总是可以减少新网段中的负载。“几乎”一词使我们不得不考

虑这样一种很少见但又是存在的现象：一个网络被分成一些网段，但是网段内部的流量为零！换句话说也就是所有的流量都是网段间的流量。图15-3就给出了这样一个例子。它表示的是S1网段中的计算机只与S2网段中的计算机交换数据，反之亦然。

实际上，在任何网络中都有可能选择一组职员的计算机执行共同的任务。这些职员可能属于同一个工作组、部门或者公司的其他结构单元。一般情况下，他们只需要访问他们部门的网络资源，而很少需要访问远程资源。

在20世纪80年代，有这样一条经验规则：把网络分成多个网段以保证80%的流量是用来访问本地资源，只有20%的流量是用来访问远程资源是有可能的。这样一条规则并不总是对的。相反，在实际中这个比例可能会变成50%~50%甚至是20%~80%。例如，有可能大多数的资源访问都是要访问互联网或者是集中在公司服务器上的资源。不过，网段之间总会有内部流量的，否则这个网络的分段方法就是不对的。

子网改进了网络的灵活性（Subnet improve network flexibility）。当把网络作为网段（子网）的集合来构建时，每一个子网都可以根据某个工作组或者部门的具体需求进行调整。例如，一个子网可以使用以太网技术和UNIX操作系统，而同时另一个子网可能是基于令牌环技术并使用OS-400操作系统，这主要是由具体部门或现有的应用需求决定的。两种子网的用户可以通过网桥或者交换机交换数据。这样，将网络分割成不同的逻辑分段的过程可以从一个相反的角度来考虑，也就是通过连接现有子网来建立大型网络。

子网加强了数据的安全性（Subnet strengthen data security）。通过在网桥或者交换机上安装多种逻辑过滤器，就可以控制用户对其他网段的资源的访问。要注意的是中继器不提供这样的能力。

子网简化了网络的管理（Subnet simplify network management）。网络管理的简化是流量减少和数据安全性加强的一个副产品。因为问题常常局限在一个网段之内。网段形成了网络管理的逻辑域。

我们已经提到过，一个网络可以通过网桥或/和交换机被分成多个逻辑网段。在20世纪90年代早期，交换机刚刚出现的时候，生产这种新设备的公司的市场部门试图给人们制造这样一个错觉，那就是网桥和交换机是不同的设备。

然而，网桥和交换机有如此多的共同点，以至于在功能上它们就像是双胞胎。网桥和交换机之间主要的不同就是前者是串行地处理帧而后者是并行地处理帧。

这两种设备都是基于相同的算法转发帧。这个算法在IEEE 802.1D标准下叫做**透明网桥算法**（transparent bridge algorithm）。

这个标准，在第一台交换机出现之前就早已被用来描述**网桥**（bridge）的操作。因此，术语**网桥**至今仍然很自然地被保留在这个算法的名称之中。当第一个交换机模型出现的时候，由于交换机的操作是基于帧转发算法（这个算法在IEEE 802.1D标准中被描述）而产生了一些混淆。网桥已经使用这个算法有将近十年了。虽然，这个算法所服务的网桥已经是过时的通信设备，在实际中已经不再使用了，但是传统上术语**网桥**仍然被标准用来描述交换机的操作。当然我们并不是如此守旧的。当在下一节描述802.1D标准中的算法的时候，除了在提及一个标准的官方名称或者需要强调这两种设备的不同时会用到网桥，我们都只使用术语**交换机**（switch）。

15.2.3 IEEE 802.1D标准的透明网桥算法

在透明网桥算法名称中的**透明**（transparent）这个词表现了这样一个事实，就是网桥和交换机在操作中不用考虑端节点的网络适配器、集线器和中继器。另一方面，上面列出的设备在工作时也不用管网桥或者交换机的存在与否。

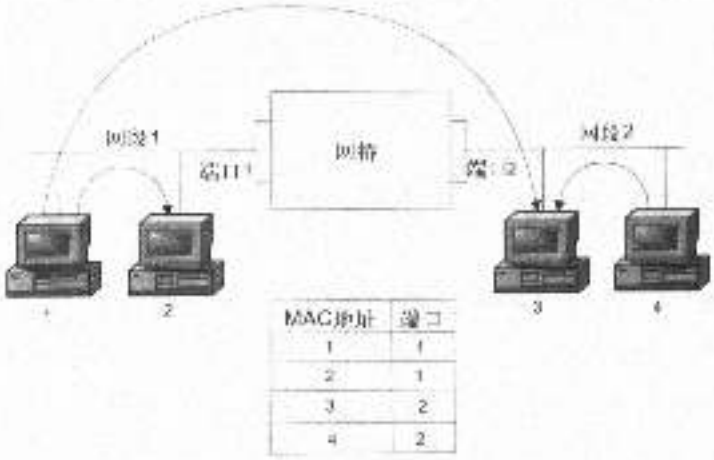
无论在安装了网桥的LAN中使用了哪一种LAN技术，透明网桥或交换机算法都成立。所以，

以太网透明网桥或交换机与FDDI或者令牌环中透明网桥或交换机的操作方法是相同的。

交换机通过被动地监察连接在其端口上的网段的流量来建立地址列表。交换机要考虑流进交换机端口的数据的源地址。交换机还要通过节点发出的帧中的源地址确定这个特定的源节点来自于哪个网络分段。

说明　交换机的每个端口都好像是一个分段的端节点，除了一个例外——即，那个交换机端口没有自己的MAC地址。交换机端口不需要地址是因为它们运行于帧捕捉的混杂模式下。在这个模式下，所有的流入端口的帧被载入缓存中，而不考虑它的目标地址。当处于这种混杂的模式下时交换机"侦探"它所连的网段内流通的所有流量，以此来学习网络的结构。

思考在图15-4这个简单的网络例子中交换机是如何自动生成和使用地址列表的。



MAC地址	端口
1	1
2	1
3	2
4	2

图15-4　透明网桥或交换机的操作原理

交换机连接两个网络分段，网段1是由一段同轴电缆所连接计算机组成的一个网段，这段同轴电缆与交换机的端口1相连。网段2是由另一段同轴电缆所连接的计算机组成的。网段2与交换机的端口2相连。

一开始，交换机不知道它每个端口所连的计算机的MAC地址，在这种情况下，交换机只是简单地将捕获并缓存的帧转发到除该帧的接收端口外的其他所有端口。在这个例子中，交换机只有两个端口，所以它将帧从端口1传送到端口2，或者将帧从端口2传送到端口1。在这种模式下的交换机与中继器的不同之处就是：交换机总是先缓存整个帧再转发，而不是1比特1比特地转发。缓存使得所有网段的逻辑操作变成了好像只有一个共享介质。当交换机将要从一个网段向另一个网段传输帧的时候——例如，从网段1向网段2——它通过使用一种特定的介质访问算法（在这个例子中为CSMA/CD）使得访问网段2就好像是访问一个终端节点一样。

在帧传送到所有端口期间，交换机学习到帧的源地址，并在交换机的地址表中添加如下记录，即这个地址属于哪个网段。这个表也被称做过滤表或者是路由表。例如，从计算机1收到发往它的端口1的帧，交换机就将这条记录记录到它的地址表中。

*MAC*地址1　端口1

这条记录表示MAC地址为地址1的计算机属于连接到交换机端口1的那个网段中。如果这个网络中的四个计算机都是活动的并相互之间交换帧，那么交换机很快就可以建立起这个网络的完全

地址表，这个地址表包含4条记录，每个节点一条记录（图15-4）。

每当一个帧到达交换机的端口时，交换机便用它的地址与所有到达帧的目的地址进行比较看它们是否有匹配。现在我们继续讨论图15-4例子中交换机的操作。

1. 收到一个从计算机1发送到计算机3的帧，交换机就查找地址表来寻找与帧中指定的目标地址所匹配的地址：MAC地址3。表中有这样的记录。

2. 交换机进行表分析的第二阶段。在第二阶段，交换机检查源地址的计算机（MAC地址1）与目的地址的计算机（MAC地址2）是否在同一个网段中——换句话说，它们是否连接在同一个端口上。在这个例子中，计算机1和计算机3是在不同的网段中，因此，交换机就执行了被称之为帧转发的操作——也就是说它将帧转发到另一个端口，这个端口事先已获准访问的另一网段。

3. 如果交换机发现帧中的源地址和目标地址的计算机属于同一个网段，交换机就只需要把这个帧从缓存中删除掉，这一操作被称为过滤。

4. 如果目标地址**对于交换机来说是未知的或未学习到的**，也就是说在地址表中找不到对应的记录，交换机就把帧传送到除了源端口以外的所有端口。这个过程与一开始的学习过程是相似的。

交换机的学习过程是永不停止的，并且它是与帧转发和过滤是并行的。交换机不断地追踪缓存中帧的源地址，以此来适应网络中不断发生的变化，例如将计算机从一个网段移至另一个网段，移除计算机或者移入新的计算机。

地址表可以是*动态的*，在交换机自学习的过程中被创建；也可以是*静态的*，由网络的管理员手工创建。**静态项**（static entry）没有终止时间，这样可以允许管理员人为地影响某个指定计算机地操作——例如，限制有特定源地址的帧从一个网段传输到另一个网段。

动态项（dynamic entry）有终止时间——当地址表中的一个已有项被更新或者建立了一个新的项时，它将会与一个时间戳相关联。预先定义的时间到后，如果交换机在这段时间内收到的帧中没有一个帧的源地址域中的地址是这条地址，那么这条记录就会被标志为无效的。这就提供了对于像计算机从一个网段转移到另一个网段的事件发生的自动反应。如果一个计算机不再继续连接到它所处的网段时，地址表中关于这台计算机属于这个网段的信息不久之后就会自动被删掉。当这台计算机连接到另一个网段时，它的帧将会通过另一个端口到达交换机的缓冲区，这样，根据网络的变化情况，一条新的记录将会被记录到地址表中。

包含有广播MAC地址的帧和包含有未知目标地址的帧将会通过交换机传送到它所有的端口。这种帧的传播模式称为*洪泛*。网络中的交换机不会阻止任何网段的广播帧的传播，因此保留了它的透明性。然而，这种特性只有在建立广播地址的节点正确运行的情况下才是有益的。

然而，情况常常是这样的，由于软件或者硬件的故障，上层的协议或者网络适配器开始不正确运行了，换句话说就是在长时间内不断地发送广播帧。在这种情况下，交换机把这些帧传送到所有的网段中，这样使得网络被错误的信息所淹没了。这被称之为*广播风暴*。

不幸的是，交换机无法保护网络不遭受广播风暴，至少在默认情况下是不行的——与路由器相比（路由器的这个性质我们将在第四部分进行讨论）。为了阻止广播风暴，管理员最多只能借助于交换机为每个节点指定发送广播帧的最大强度。此时，我们又必须精确地知道什么强度是正常的，什么强度表示出错。当改变协议时，网络状况可能发生改变。特别地，有可能这种状况在昨天还被认为是错误的，今天又被证实为非常正常。

图15-5展示了交换机的一个典型结构。当接收和传送帧时，介质访问功能是由MAC电路执行的，这类似于网络适配器。

图15-6展示了实现交换机算法的协议是在MAC和LLC层之间。

图15-7展示了一份本地交换机模式下终端显示的地址表。终端连接着控制台端口，显示在屏幕上的信息是由交换机控制单元生成的。

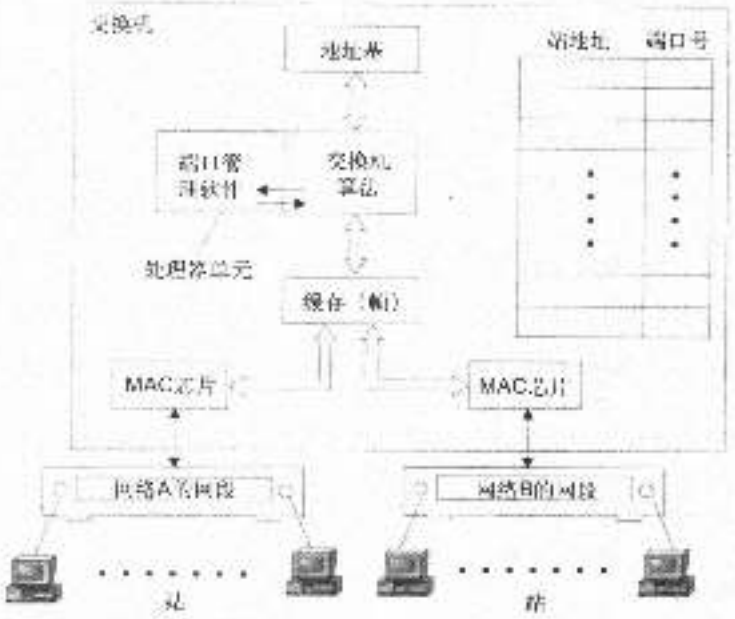


图15-5　交换机结构



图15-6　协议栈中交换机协议的位置



图15-7　交换机的地址表

从屏幕上显示的地址表（转发表）中，我们可以看到这个网络由两个网段组成——LAN *A*和LAN *B*。在LAN *A*网段中，至少有三个站，在LAN *B*网段中，有两个站。四个有星号标识的地址是静态地址（即它们是由网络管理员手工分配的）。由加号标识的地址是动态地址，是有终止时间的。

表中有一列，名为*Disp*——处理。这一列的数据告诉交换机需要对某特定目标地址的帧进行

什么操作。当这个表被自动创建时，这个区域常常包含目标端口的常规名称。然而，当地址是手工指定时，可以在这个区域指定非标准的硬过程操作。例如，洪泛操作使得交换机在广播模式下发送帧，即使它的目标地址不是一个广播地址。丢弃（Discard）操作让交换机在遇到指定的地址时不是将其转发到指定的端口，而是将其丢弃。

*Dispn*列中定义的操作指定了帧过滤的特定的条件它们传播的补充标准条件，这些条件常常被称为用户自定义的过滤器。我们将在15.3.5节继续讨论。

15.2.4 交换机LAN的拓扑局限性

这些通信设备的一个严重的局限就是它们不能支持网络的环路配置。

考虑图15-8例子中的网络设备的局限性。

在这个例子中，两个以太网段同时由两个交换机并行连接从而形成了一个活动的环路。假设一个MAC地址为123的新的站点是第一次连接到这个网络中并且开始操作。通常，启动任何一个操作系统过程都是伴随着传送广播帧，在这些广播帧中站点通知其他计算机自己已经连接到网络中同时查找网络服务器。

第一步，站点发送第一个帧到本地网段中，这个帧的目标地址为广播地址，并且源地址为123。这个帧首先到达交换机1和交换机2。在两个交换机中，这个新的源地址，123，被写入到地址表中，并且标记为属于网段1。这意味着建立了一条新的地址表记录，如下所示：

MAC地址	端口
123	1

因为帧的目标地址是广播地址，所以每一个交换机都必须把这个帧传送到网段2。这个传输根据以太网技术的随机访问方法依次进行。假设交换机1是第一次获准访问网段2（图15-8中的第二步）。当帧到达网段2时，交换机2接收它，并把它加载到它的缓冲区中，然后对它进行处理。交换机2注意到地址123已经在它的地址表中。然而，刚刚到达的帧更新一些，而且它说明地址123属于网段2而不是网段1。因此，交换机2修改了地址表的内容，建立了一条新的记录指定地址123属于网段2。

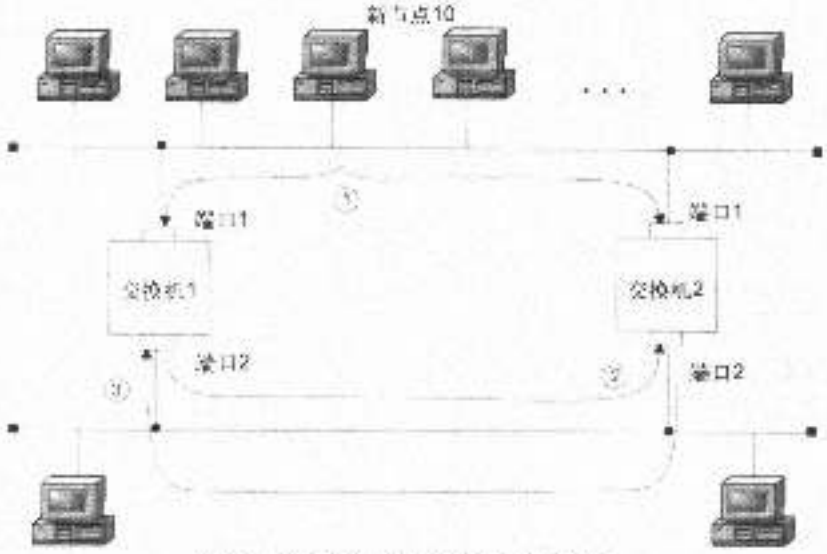


图15-8 闭合路由对于交换机操作的影响

MAC地址	端口
123	2

当交换机2把该帧的副本传送到网段2中时，交换机1以相同的方法处理。

因此，环路导致了以下的后果：

- 帧在“生殖”（例如，出现了同一个帧的多个不同拷贝）。在上面这种情况下，出现了两个拷贝；如果网段由三个交换机连接，就会有三个拷贝，等等；
- 两个帧的拷贝在相反方向上沿着环路无止境地循环，这将导致网络充满了没有用的流量；
- 交换机地址表不断地重构，这是因为源地址为123的帧经常在两个端口交替出现。

为了消除所有的这些不需要的影响，交换机必须能够消除逻辑网段间的环路。这表示交换机只允许建立树状的网络结构，这种结构能保证任何两个网段之间只有一条路由。每个站点的帧只能到达交换机的同一个端口，并且交换机可以正确地在网络中选择一个合理的路由。这就是树形网络拓扑。

在小型网络中，在两个网段中保证只有一条可能的路径存在是相对容易的。然而随着链路数的增长，无意地创造环路的可能性变得越来越高。

此外，为了增加网络的可靠性，需要交换机间的预留链路。在主要链路操作正常的情况下，它们是不参与帧传输的。但是当主要链路出现故障时，就需要在不产生环路的情况下使用预留链路修复连通性。

因此，在复杂的网络结构中，在网段间建立冗余的链路。这些冗余的链路会形成环路。为了消除这些活动环路，交换机的一些端口必须被锁定。解决这些问题最简单的方法就是手动配置。然而，存在自动解决这些问题的算法。其中最有名的就是**生成树算法**（spanning tree algorithm, STA），我们将在第16章讨论它的细节。

15.3　交换机

15.3.1　交换机的特殊性质

20世纪80年代后期到20世纪90年代早期发生的那场根本性的网络革命不但将网络划分成大量的网段，还导致了快速协议、高性能的PC机和多介质信息技术的到来——传统的**网桥**已经不再能完成这些任务。若使用单一处理单元服务多个端口间的帧流，则对处理器运行速度的要求有很大提高，同时这也是一个昂贵的解决方案。

一个更有效的解决方案是：为了处理到达每一个端口的数据流，为设备上装备了单独的使用网桥算法的处理器。这个解决方案也促进了交换机的发展。据其特征，交换机是一个多处理器的网桥，可以同时在所有成对的端口间转发帧。然而，与计算机不同的是，当计算机添加了新的处理器后它们的名称并不改变，仅仅变为“多处理器配置”，而多处理器网桥则被称为交换机。设备名称的改变，是由组织交换机内处理器间连接的方法引起的——它们由一个交换矩阵连接，这个矩阵与多处理器计算机连接处理器与内存的矩阵类似。

渐渐地，交换机把老式的单处理器网桥挤出了LAN。主要原因就是交换机保证了在不同的网络分段间传输帧时地高性能。与可能会减缓网络操作的网桥相比，交换机通常都装备有能够以协议所能允许的最大速度传送帧的端口处理器。加上可以在多端口之间并行传输帧的功能使得交换机要比网桥的性能高上10多倍。下面这个事实展示了网桥和交换机之间的发展前景。

交换机可以每秒传输几百万个帧；而网桥每秒只能传输3 000至5 000个帧。

在网桥和交换机同时存在的时候，没有网桥的竞争，交换机采纳了很多作为网络技术发展的必然结果的额外功能。这些功能包括支持虚拟局域网（VLAN）、流量优先级划分以及使用默认的主干端口。

1990年，一家叫做Kalpana的小公司首先提出了交换以太网分段技术。这个技术的提出是为了解决对于连接高性能服务器以及包含工作站的网段之间日益增长的带宽要求。

如果输出端口在帧接收时是可用的，则对于Kalpana的交换机而言，从接收到第一个字节至该字节被送到输出端口，其时间延迟仅仅为40μsec。相对于由网桥传输帧的延迟相比，这是一个非常明显的优势。

图15-9展示了Kalpana提出的EtherSwitch交换机的结构。

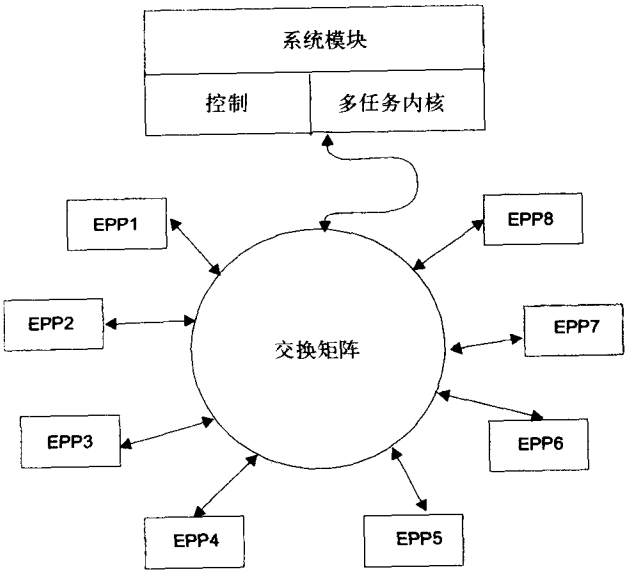


图15-9 Kalpana提出的EtherSwitch交换机的结构设计

8个10Base-T端口中的每一个都由一个**以太网分组处理器**（Ethernet packed processor, EPP）来处理。不仅如此，交换机还有一个系统单元用来协调所有EPP的操作。系统单元提供交换机的共享地址表。交换矩阵用来在不同的端口间传输帧。它根据电路交换原则运行并连接交换机端口。对于8个端口，当传送方和接收方互相独立时，这个矩阵可以在使用半双工操作模式时保证8个同时存在的内部信道；在使用全双工模式时保证16个信道。

当一个帧到达其中一个端口时，EPP缓存帧的头几个字节来读取目的地址，接收到目的地址后，处理器不用等待该帧的剩余字节就可以立刻决定帧的传输。为了达到这个目的，它查看自己的快速缓存中的地址表。如果没有找到需要匹配的地址，EPP就转换系统模式为多任务模式，并行地处理所有的EPP请求。系统模块查找公共地址表并返回所需要的条目给处理器。EPP同时缓存这条信息以备后用。

- 如果在地址表中可以找到目的地址，并且新到的帧必须被过滤掉，那么处理器所要做的仅仅是停止将帧载入缓冲区，清空缓冲区，并且等待新帧的到来。
- 如果在地址表中可以找到目的地址，并且新到的帧必须被转发到另一个端口，那么处理器一边对交换矩阵编址，一边继续将帧载入到缓冲区中，并且尝试建立连接本端口与到目的地址需要经由的端口的路径。交换矩阵只有在目的地址端口空闲时才能这样做，也就是它目前不能与其他端口相连。
- 如果端口正忙，矩阵拒绝了连接请求，这种情况对于任意基于电路交换的设备都是可能的。在这种情况下，帧被输入端口由处理器完全载入了缓冲区，然后处理器等待直到输出端口被释放并且交换矩阵可以建立需要的路径。

- 建立好需要的路径之后，在缓冲区中的帧字节就通过它送到输出端口的处理器。一旦输出端口的处理器获准访问它所连接的以太网段（使用CSMA/CD算法），帧就立刻开始传送到网络中。输入端口的处理器不断将收到的帧的一些字节存储在缓冲区中，这使得它可以独立地并且异步地传输和接收帧字节（图15-10）。

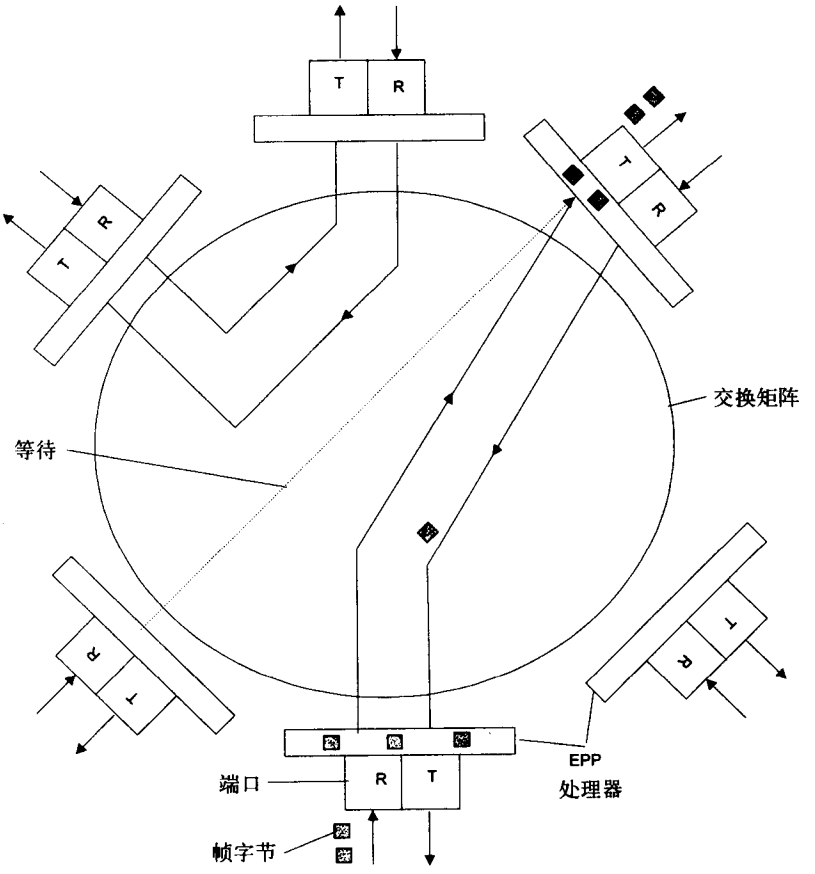


图15-10　使用交换矩阵的帧传输

这种无需完全缓存的帧传输方法被称为直通交换或者切入通交换。原则上，这种是管道帧传输过程，其中的一些传输步骤是并发进行的。这些步骤包括：

1. 输入端口的处理器接收帧的头几个字节，包括那些含有目的地址的字节。
2. 在交换机的地址表中、或者在EEP高速缓存中、或者在系统模块下的公共表中寻找目的地址。
3. 交换矩阵。
4. 输入端口处理器接收帧的余下字节。
5. 输出端口处理器通过交换矩阵接收帧字节（包括头几个）。
6. 通过输出端口处理器访问介质。
7. 通过输出端口处理器将帧字节传输到网络中。

图15-11列举了两种帧处理模式：多个传输步骤是并行执行的管道帧处理，和完全缓冲并且顺序执行每一步的普通帧处理。（注意步骤2和步骤3不能并行执行，因为如果不知道输出端口号，交换矩阵操作就没有任何意义）。

正如从上面这个例子所看到的，与完全缓冲模式相比，这种通过管道而节省的时间相当的可观。

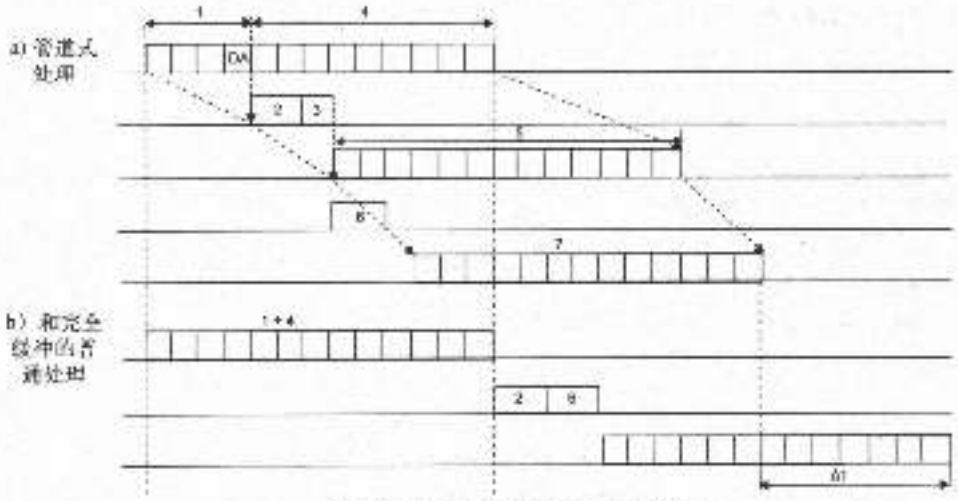


图15-11 管道式的帧处理节省的时间

然而，使用交换机可以提升网络性能的最主要的原因就是交换机可以并行处理多个帧。

图15-12向我们阐释了这个作用。一个理想的性能：8个端口中的4个正在以以太网协议所允许的最大速率——10Mb/s——传送数据，并且这些数据是被传送到余下的4个端口上，但是没有产生冲突。不存在冲突的意味着数据在网络节点间的传输是分布式的，这样每个收到帧的输入端口都有一个可用的输出端口。如果在输入端口的帧强度达到的情况下，交换机仍可以成功地处理输入流量，则这个例子中交换机的总性能将为4×10=40Mb/s。如果将这个例子中的端口归纳为N个，则交换机的总性能将是（$N/2$）×10Mb/s。**在这种情况下，交换机给每个连接在其端口上的站或者网段提供协议所分配的带宽。**

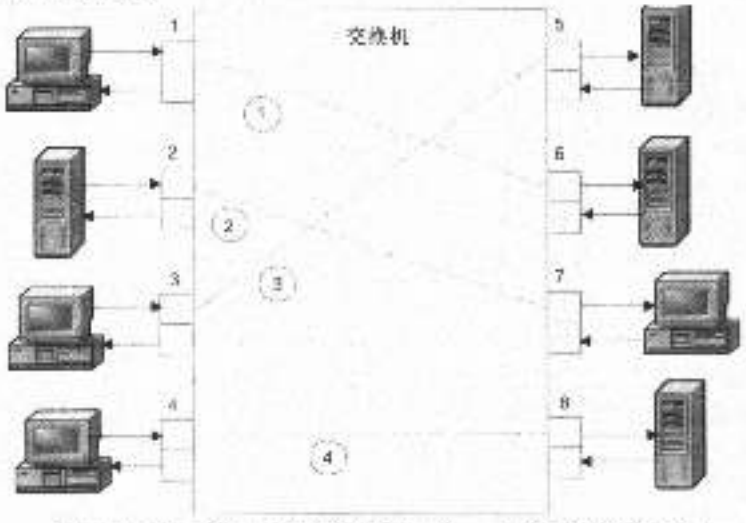


图15-12 通过交换机并行传输帧（1~4——计算机之间的帧流）

无疑，这种情况在网络中并不是经常发生的。例如两个站（假设它们是连接在端口3和端口4上）需要同时向同一个服务器写数据，这个服务器连接在端口8上。在这种情况下，交换机将不能对每个站都分配10Mb/s的数据流，因为端口8不能以20Mb/s的速率传输数据。两个站的帧要在端口3和端口4的内部队列中等待直到端口8可以用于传送下一帧。对于这样的数据流分布，只有在服务

器连接到如快速以太网这样的速度更快的端口时才有意义。

15.3.2　无阻塞的交换机

无阻塞的交换机是这样一种交换机：它可以以帧到达端口的速率传送帧通过端口。

通常，当说到交换机操作的稳定的非阻塞模式时，我们假设交换机在任意时段内都以帧到达时的速率传送帧。为了保证这种操作模式，就必须实现帧流的这样一种分布：在这种分布下输出端口可以成功地处理负载。如果观察到这种要求，交换机就可以在一定数目的帧到达输入端口时总是传送平均来说相同个数的帧给输出端口。如果帧输入流（所有端口的总和）平均来说超过了帧的输出流（仍然是所有端口的总和），这些帧就会累积在交换机的缓存中。如果超出可用的缓存（也就是缓存溢出），交换机就开始丢弃帧。

为了保证交换机操作的非阻塞模式，必须满足下面这个简单条件：

$$C_k = (\Sigma C_{pi})/2 \tag{15.1}$$

这里，C_k是交换机的性能，C_{pi}是交换机的第i个端口所支持的协议最大性能。

端口的总性能对每个通过的帧都考虑了两次——第一次是作为输入帧，第二次是作为输出帧。由于在稳定的操作模式下，输入流量与输出流量是相等的，为了支持非阻塞模式，交换机最少应达到的性能应当是它所有端口的总性能的一半。如果端口在半双工模式下，例如10Mb/s以太网，那么端口的性能（C_{pi}）就是10Mb/s；如果它工作在全双工模式下，它的性能就是20Mb/s。

有时候，声明宣称交换机可以保证即时的非阻塞模式。这就表示交换机可以从它所有的端口以支持协议所保证的最大速率接收和发送帧，无论输入和输出之间稳定且平衡的流量是否达到。说实话，有一些帧可能处理的不完全——如果输出端口忙碌，帧就会被载入到交换机缓冲区中。

为了支持操作的即时的非阻塞模式，交换机必须支持更高的总性能，在这种情况下必须与所有的端口的性能之和相等：

$$C_k = \Sigma C_{pi} \tag{15.2}$$

第一个LAN交换机被设计成为以太网技术服务并不是一个偶然。除了以太网极其普及之外，还有一个非常重要的原因就是：由于当网段超载的时候，它必须等待直至可以访问介质，这个性质使得这个技术非常易于受延迟增长的危害。正因如此，在大型网络中的以太网网段就第一个需要消除网络瓶颈。Kalpana交换机以及后来来自于其他厂家的交换机正式提供了解决这个问题的方法。

有一些公司开始致力于开发提高其他局域网技术性能的交换技术，这些局域网技术包括令牌环和FDDI。来自于不同生产厂家的交换机其内部结构往往与最初的EtherSwitch交换机结构有着显著的不同；然而，通过每一个端口进行的帧并行处理的原则并没有任何的变化。

刺激交换机被广泛使用的原因是引入交换技术并不需要更换包括中继器、集线器以及电缆系统等一些已安装的网络设备。交换机端口工作于普通的半双工模式，因此它们允许端节点和组织起整个逻辑网段的集线器之间的透明连接。

因为交换机和网桥对于网络层协议是透明的，所以网络中如果存在路由器的话，它们的引入也不会有任何影响。

15.3.3　克服拥塞

在典型的半双工模式下，交换机具有使用介质访问算法机制影响端节点的可能性，这个机制是它的邻居节点所必须执行的。主要有两种用于控制帧流的方法：作用于端节点上的背压技术和主动介质抢占。

背压技术是在网段中以过高的强度发送帧给交换机，来创建人工的拥塞。为了这个目的，交

换机通常都会使用阻塞序列，将它发送到连接在交换机输出端口的网段上（或者节点），从而对它的活动进行挂起。

当邻居节点是端节点时，通常采用第二种减缓帧流的办法。这个方法是基于当抢占介质时交换机端口的主动行为，或者在传输下一帧之后，或者在冲突之后。图15-13列出了这两种情况。

在第一种情况下（图15-13a），交换机完成了下一帧的传输。它只停止了9.1μsec而不是技术上所要求的9.6μsec，就开始传输一个新的帧。计算机无法获取介质，因为它持续等了一个标准停顿时间也就是9.6μsec，所以仍然发现介质还是处于忙的状态。

在第二种情况下（图15-13b），交换机的帧和计算机发生冲突，并且这个冲突被登记下来。因为计算机在冲突后按标准的要求暂停了51.2μsec（延迟的间隙时间是512比特），且交换机暂停了50μsec，所以计算机仍然无法传输它的帧。

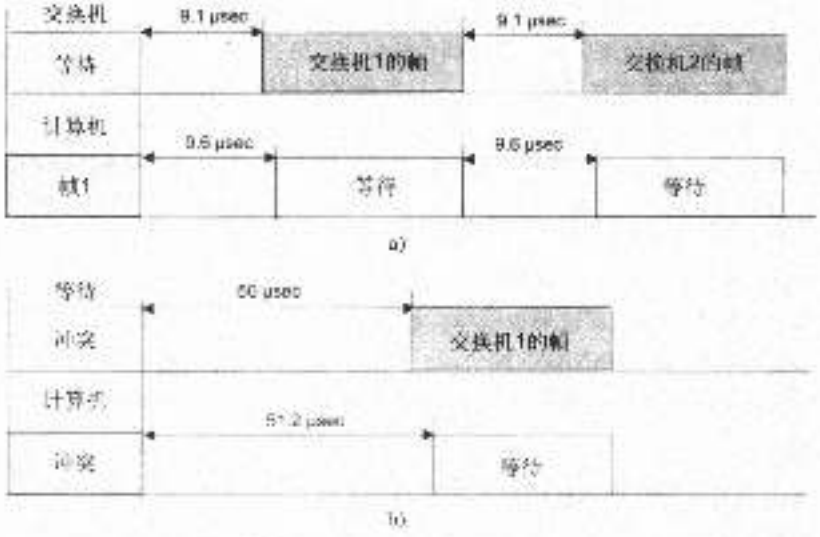


图15-13 缓冲区溢出的情况下交换机的主动行为

交换机可以根据情况使用这个机制，如果需要的话可以增加它的主动级别。

许多生产商使用相当复杂的机制，也就是把这两种拥塞控制方法合二为一。这些方法使用基于交互传送和接收域的帧交错算法。帧交错算法必须要有足够的灵活性以让交换机在临界情况下每接收到一个帧就传送多个帧，这样可以释放内部的帧缓冲——不用将帧接收速度减少到0，而只要减少到需要的层次上。

15.3.4 数据链路层协议的翻译

交换机可以通过IEEE 802.1H标准和RFC 1042将数据链路层协议转换为其他协议，例如，以太网变成FDDI或者快速以太网变成令牌环。

LAN协议转换是很简单的，因为最难的工作——地址转换，是由连接在不同种类的网络中的路由器和网关来完成——在这里是不需要的。

LAN中的所有的节点都有一个格式相同（MAC地址）的唯一地址，并且独立于支持它们的协议。

因此，FDDI网络的网络适配器可以读懂以太网的网络适配器地址，并且两个节点都可以在它们的帧中使用这些地址而不用担心与它所交互的目的节点属于运行在不同技术下的网络中。

因此，当调整LAN协议的时候，交换机不用建立地址转换表，相反，它们可以在不同的协议帧之间传递帧的源地址和目的地址。

除了在传输地址字节时改变比特顺序，从以太网协议（以及使用相同帧格式的快速以太网协议）到FDDI和令牌环协议的地址转变还执行下面的一些或者全部的操作：

- 在从FDDI或者令牌环网络传输帧到以太网802.3网络的时候，计算帧的数据域的长度，并且将这个值放进帧的长度域中。（FDDI和令牌环网络不包括长度域）。
- 在从FDDI或者令牌环网络向以太网网络传输帧时，必须填充帧状态域。FDDI和令牌环网络的帧中有两个比特用来标识帧的状态：地址识别比特（A）和帧拷贝比特（C）。当交换机将帧传输到另一个网络又通过环返回到源端时，没有标准的规则用来置帧中的A和C的位。因此，交换机的生产者们根据自己的判断来解决这个问题。
- 当从FDDI或者令牌环传送帧到以太网时，以太网会抛弃那些数据域超过1 500比特的帧，因为1 500比特的数据域是以太网中帧的数据域的最大允许值。然后，位于FDDI或令牌环网络中的源站发现没有收到以太网络的回复，它们上层的协议就很可能会减少在一个帧中传输的数据。之后，交换机就可以在这些站间传输帧了。另一种解决方法是确保交换机支持IP分段。可是，这要求交换机实现网络层协议，并且确保转换网络的交互节点支持IP协议。
- 当传送来自于支持FDDI或者令牌环网络的帧时，填充以太网Ⅱ帧中对应的类型域（数据域中协议的类型），因为它们的帧中没有这个域。FDDI或者令牌环帧有DSAP和SSAP域，与以太网中的类型域的目的相同，只是用其他编码来指定协议。为了简化这种转换，FDDI和令牌环网络常用的RFC1024规范建议帧使用LLC/SNAP的帧头，它们与以太网Ⅱ的帧有着相同的类型域且值也相同。当转换帧时，来自于LLC/SNAP的帧头中的类型域可以直接转移到以太网Ⅱ中帧的类型域，反过来也一样。如果有与在以太网中的以太网Ⅱ格式不同的帧格式，则它们也必须使用LLC/SNAP的帧头。
- 通过帧服务域新形成的值重新计算帧的校验和。

15.3.5　流量过滤

许多交换机模型允许管理员指定额外的帧过滤条件，用来作为通过地址表信息生成的标准帧过滤条件的补充。

用户自定义过滤器（User-defined filter）是用来给帧创造额外的屏障，从而达到限制某些用户访问特定的网络服务的目的。

最简单的用户自定义过滤器建立在站的MAC地址基础上。由于MAC地址是与交换机运行有关的信息，所以它可以允许网络管理员很方便地建立这种形式的过滤器。例如地址表中可以指定一个额外的域用于存放一些条件，就如同图15-7中显示的交换机地址表中指定那个（例如丢弃特定地址的帧）。这样，使用拥有特定MAC地址的计算机用户就不能访问另一个网段的资源。

常常，管理员需要指定一些更复杂的过滤条件，例如，可以让其他网段的用户访问这个网段的所有其他资源，但是不能从另一个网段打印某个Windows打印服务器上的文件。为了实现这样的一个过滤器，阻止特定MAC地址的帧传输是必须的。这些帧可能包括封装的SMB分组，如果这个分组的对应域里指定了“打印”服务类型。交换机不分析这些上层协议，例如SMB；因此，管理员为了指定过滤条件，必须手工地确定需要过滤的域值。这个过滤器以相对于数据链路层帧的数据域的开始位置的“偏移量大小”对的形式被指定，然后就必须指定与这个打印服务有关的16进制值。

通常，过滤条件被写成布尔表达式，由AND和OR操作符组成。

15.3.6　交换机体系结构和设计

如今，为了加速交换机的操作，所有的交换机都使用专用的LIC：最适宜执行交换机操作的ASIC。通常，一个交换机有多个专用的LIC，每一个都执行一组功能完备的操作。

除了一些特殊的处理器芯片外，在非阻塞模式下的成功操作还要求交换机装备有一个快速单元，这个快速单元用于在交换机端口的处理器芯片间传输帧。

目前，交换机使用以下三个基本方法之一来建立交换单元：

- 交换矩阵
- 公共总线
- 共享多端口内存

常常，这三种方法被结合到同一个交换机中。

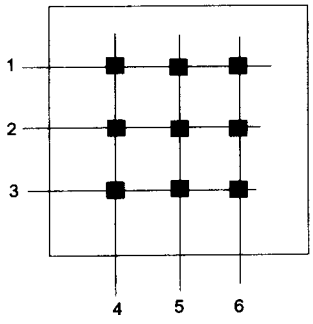


图15-14　交换矩阵

交换矩阵（switching matrix）保证了使用最简单的方法在不同的处理器端口间的交互。这种方法在第一批LAN交换机上就被实现了。然而，交换矩阵只有在端口数目有限且已事先定义好的条件下才有可能实现，同时设计的复杂度与交换机端口数量的平方成比例（图15-14）。

图15-15展示的是另一个支持8个端口的交换矩阵的实现。查找交换机的地址表得到端口处理器的输入，根据目的地址判定输出端口的编号。它们将这些信息以特殊标签的形式加在源帧之上。在这个例子中，为了简化起见，标签是与输出端口号对应的3比特的二进制数字。

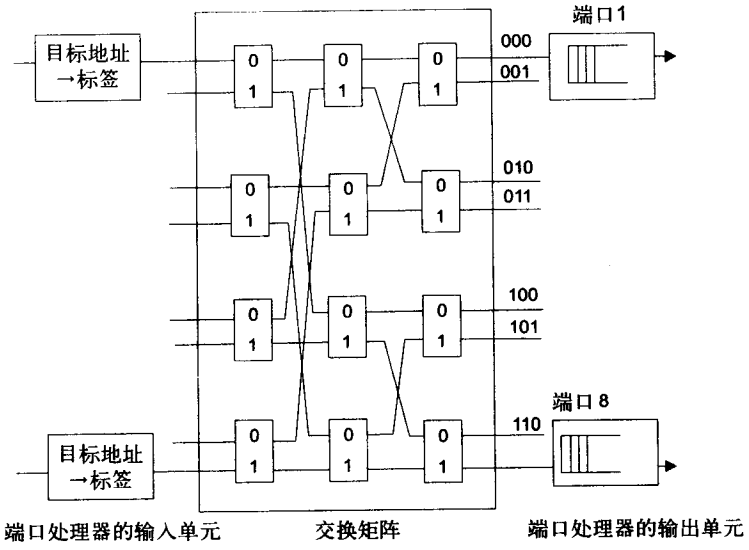


图15-15　使用二进制交换机的8×8交换矩阵的实现

矩阵是由三层二进制交换机组成，这些交换机根据标签的位值连接它的输入到两个输出中的一个。第一层交换机由标签的第一个比特来控制，第二层交换机由第二个比特来控制，第三层交换机由第三个比特来控制。

这个矩阵可以用基于其他组合设计的其他形式的方法来实现。然而，交换物理链路的技术仍然保持它所特有的特点。这个技术有一个众所周知的缺点，那就是交换矩阵缺少数据缓冲。因此，如果由于中间交换机的某个或多个元件的输出端，忙而导致的电路无法形成时，数据就必须在数据源端累积，在这个例子中数据源端的角色是由接收帧的端口的输入单元扮演的。这种矩阵的主要优点包括高速的交换和规则的结构，这在LIC中实现起来很方便。然而，在实现$N*N$矩阵作为LIC的一部份之后，另一个缺陷变得非常明显，即很难增加交换机端口的数目。

在基于**公共总线**（common bus）的交换机中，端口处理器是用于分时模式下的高速公共总线连接的。

图15-16展示了这个结构的一个例子。为了确定总线不会阻塞交换机的操作，必须保证它的性能至少是该交换机所有端口的性能总和。对于模块化交换机，低速端口的模块结合可以实现非阻塞操作，而安装高速端口模块的可能会导致公共总线变成瓶颈。

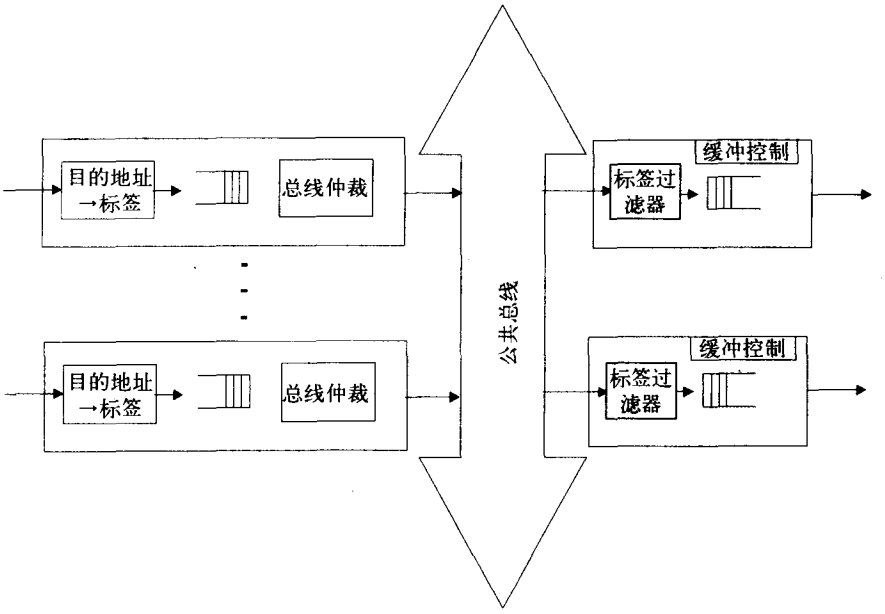


图15-16　基于公共总线的交换机结构

帧必须以多个字节组成的小部分通过总线来传送，使得端口之间的数据传输在伪并行模式下进行，且在整个帧的传输过程没有引入延迟。这样的数据信元的大小由交换机的制造商决定。一些制造商选择选择一种ATM信元作为每次操作通过总线传送的数据部分，这种ATM信元中的数据域长度为48字节。如果交换机支持这些技术，那么这种方式就简化了将LAN协议转换为ATM协议的过程。

处理器的输入单元给这个通过总线携带的信元添加了一个标签，在这个标签里它指定了目的端口号。每一个端口处理器的输出单元都包含有为这个端口选择标签的标签过滤器。

和交换矩阵一样，总线无法执行中间缓存；然而，因为帧数据被分裂成小信元，所以这种方法没有与最初等待输出端口可用性相关的延迟——这种分组交换技术在这里代替了电路交换技术。

第三种端口交互的基本结构就是**共享内存**（shared memory）。图15-17提供了这样的一个例子。

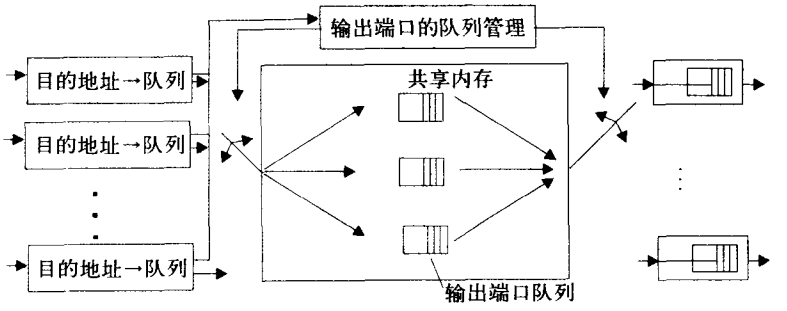


图15-17　基于共享内存的交换机结构

处理器端口的输入单元连接在交换机共享内存的输入部分，并且相同处理器的输出单元连接在交换机共享内存的输出部分。输出端口的队列管理者控制共享内存的输入和输出交换。在共享

内存内，这个管理器组织多个数据队列，每一个队列对映一个输出端口。处理器的输入单元向管理器传递请求，就是向帧中目的地址对映的端口队列中写入数据。队列管理器轮流将内存的输入与一个处理器输入单元相连，并且这个单元向特定的输出端口对映的队列中写入部分帧数据。当队列满了的时候，管理器轮流将共享内存的输出与处理器端口的输出单元相连，这样，数据就从队列中被写入到处理器的输出缓冲区中。

使用由不同端口间的管理器灵活地分配的共享缓冲内存，减少了对于处理器端口缓冲内存大小的要求。然而，内存必须足够地快以至于可以支持交换机*N*个端口间的数据传送速度。

组合交换机

我们前面讨论的每一种结构都有其优势与不足；因此，在成熟的交换机中会同时包括这三种结构。图15-18给我们展示了一种组合的结构。

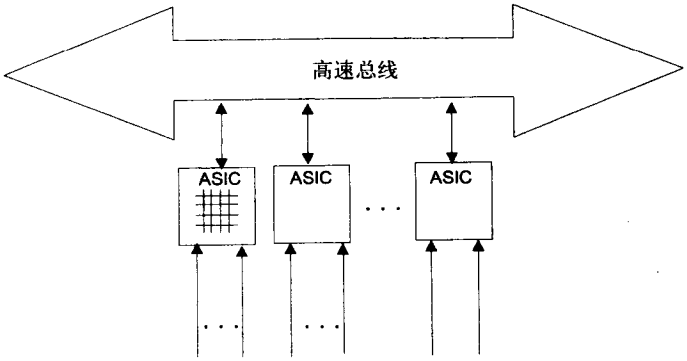


图15-18　基于交换矩阵和公共总线的组合交换机结构

交换机包含固定数目端口（2-12）的模块，是在实现交换矩阵结构的特殊LIC的基础上实现的。如果端口之间需要传送数据帧的端口属于同一个模块，那么帧的传输是由模块处理器基于该模块的交换矩阵实现的。如果端口属于不同的模块，那么处理器使用共享总线通信。有了这样的结构，由于交换矩阵对于端口间的交互提供了最快的方法，所以模块内的数据传输要比模块间的数据传输快得多，虽然这种方法的可扩展性是最差的。交换机内部总线的速率可以达到几吉比特每秒，并且一些强大的模块具有几十吉比特每秒的总线传输速率。

其他组合结构的方法也是可能的，例如，共享内存可以用来组织模块间的交互。

15.3.7　交换机的性能特性

过滤速率和转发速率是交换机的两个最重要的性能特性。这些特性是一种积分参数，因为它们并不依赖于交换机的技术实现。

*过滤速率*定义了交换机执行下面的帧处理步骤的速率：

1）将帧加载到内部缓冲区上。

2）在地址表上查找帧目的地址对应的端口。

3）如果目的地址和源地址属于同一个逻辑网段，那么就丢弃这个帧。

由于交换机有时间以帧到达的速率丢弃帧，所以过滤速率在实践中对所有的交换机来说都是非阻塞的。

*转发速率*就是交换机执行下面的帧处理步骤的速率：

1）将帧加载到内部缓冲区上。

2）在地址表上查找帧目的地址对应的端口。

3）使用在地址表上找到的端口来将帧转发到网络中。

过滤速率和转发速率通常以帧每秒来度量。如果交换机的特性中没有为转发速率和过滤速率指定协议和帧大小，那么这些参数就取默认的以太网协议和帧的最小大小——例如64字节。在相同的用户数据传输速率下，与所有具有其他格式的帧相比，具有最小长度的帧往往导致最难的操作模式。正因为如此，当测试交换机时，传输最小长度的帧的模式常常被用作最复杂的测试，以此在最差的流量参数组合下对交换机的能力进行额外地测试。

帧传输延迟是从帧的第一个比特到达交换机的输入端口，到到达交换机的输出端口的时间间隔。帧传输延迟是由帧缓冲和交换机对帧进行操作的时间组成。其中交换机对帧进行的操作包括查找地址表，决定是转发还是过滤掉这个帧，以及访问输出端口的介质。

交换机性能就是每个时间单元通过其端口的用户数据量（由兆位/秒来度量）。由于交换机在数据链路层进行操作，从它的角度来看用户数据就是在帧的数据域中的数据，并且使用的是数据链路层的协议：以太网、令牌环、FDDI等等。因为在帧头上的共享服务信息与帧的最小长度相比还要小得多，交换机性能的最大值往往在帧的长度最长时达到。由于交换机是多端口设备，因此，常常可以将同时对交换机所有端口进行信息传输的总性能作为交换机的主要性能特性。

帧传输的方法——直通或者是完全缓存传输——影响着交换机的性能。在每一个传输过程中，交换机的直通传输方法带来了最少的帧传输延迟。因此，总的数据传输延迟的减少会变得非常明显，这在多媒体信息传输中是非常重要的。此外，选择的交换方法还影响着实现一些辅助的并且有用的功能的可能性，例如，转换数据链路层协议。

表15-1提供了这两种交换方法的一个比较。

表15-1　直通交换和完全缓存交换的功能能力

功　能	直通交换	完全缓存
保护帧不出错	否	是
支持异构网络	否	是
（以太网、令牌环、FDDI以及ATM）帧传输延迟	轻负载的情况下低（5~40μsec），重负载的情况下居中	任何负载情况下都居中
支持备用链路	否	是
流量分析功能	否	是

交换机在重负载的情况下，使用直通传输产生的平均延迟是由于输出端口常常忙于接收其他帧；因此，刚刚到达当前端口的帧就必须被缓冲。

使用直通传输操作的交换机可以检查被传输帧的正确性，但是它不能抛弃错误帧，因为帧的部分字节（通常，是大部分）已经被传输到网络中了。

说明　因为每一种方法都有它的长处和缺陷，所以不允许转换协议的交换机模型有时可以实现适应性地改变交换机操作模式这个方法。虽然，这种交换机的主要操作模式是直通传输，但是交换机还要控制流量。当错误帧变得越来越频繁，并且这个强度已经超过了一个特定的阈值时，交换机就会把它的模式改变为完全缓冲传输。不久，交换机又可以返回到直通传输。

所有交换机的另一个重要的设计特性就是**地址表的最大大小**（maximum size of the address table）。它定义了交换机可以同时工作的最大MAC地址数

常常，交换机使用专用的处理器单元来执行每一个端口的操作，并且每一个处理器单元还有自己的内存来存储自己的地址表拷贝。每一个端口只存储最近使用的那些数据，因此，不同的处理器单元的地址表拷贝常常包含不同的地址信息。

处理器端口内存可以存储的MAC地址的最大数目依赖于交换机的应用领域。工作组交换机常常支持每个端口多个地址，因为它们常用来生成微网段。部门级的交换机要支持上百个地址，而主干交换机要支持上千个地址。通常是4 000~8 000。

交换机地址表大小的不足会减慢交换机的操作，并且让网络中充满了不必要的流量。如果端口处理器的地址表被填满了，而交换机又在刚刚到达的帧中遇到了一个新的地址，处理器就必须将地址表中一条老的地址替换为这个新的地址。这个操作要花费一些处理器时间。然而，当到达帧的目标地址为一个已被删除的老地址时，将要花费更多的处理器时间，因为帧中的目的地址是未知的，交换机必须将这个帧传输到给所有其他端口。

一些交换机的制造商通过改变处理未知目的地址帧的算法来解决这个问题。交换机的其中一个端口被设置为主干端口，默认情况下[⊖]，所有具有未知地址的帧都被送到这个端口。

在一个大型层次结构的网络中，如果这个端口连接的主干端口具有更高的层次和足够大小的地址表时，那么帧应该被传送到这个主干端口中。

15.4 全双工LAN协议

15.4.1 在全双工模式运行中引入MAC层的变化

交换技术并不是与交换机端口所使用的介质访问技术直接相关的。当一个表示共享介质的网段连接上一个交换机端口时，如同这个网段的任何其他节点一样，这个端口必须支持半双工模式。

然而，当交换机的每个端口连接的不是整个网段而是只有一台计算机，并且这种连接使用了两个物理上分离的信道时，除了同轴以太网外的其他以太网标准中都是切实可行的。这种情况就不再是这么确定。此时这个端口既可以以普通的半双工模式运行，也可以以全双工模式运行。

不是通过网段，而是直接连接独立的计算机与交换机端口，这就是微网段。

在以太网中常见的半双工模式操作下，交换机端口不断地侦查冲突。在这种情况下，冲突域就是包括交换机发送端、交换机接收端、计算机的网络适配器的发送端以及连接两组发送端和接收端的两根双绞线的一个网络段。(图15-19)

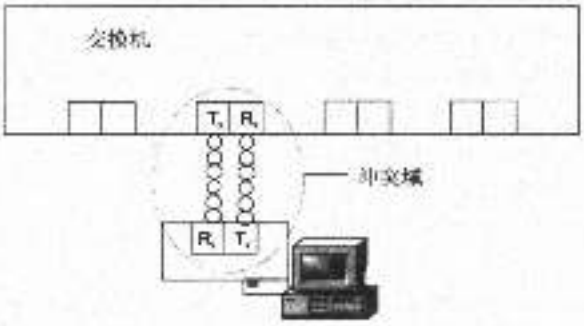


图15-19 由交换机和计算机端口组成的冲突域

假设这个网段开始是空闲的(见图15-19)，当交换机端口的发送端和网络适配器同时开始传输它们的帧时，冲突就发生了。虽然，在这样的一个分段中发生冲突的概率要明显小于由20至30个节点组成的网段，但是这个概率并不为0。同时，一个以太网段最大性能是14 880帧每秒，但是这种帧必须是最小长度的并且由交换机端口的发送端和网络适配器的发送端所共享。假设它们被分为相等的两部分，那么每一个发送端就有每秒传输约为7 440个帧的可能性。

在全双工模式下，交换机端口和网络适配器同时传输帧并不会发生冲突。原则上，在这种模式下的操作是需要独立的全双工通信信道的。它常用于WAN协议中。在全双工连接的情况下，10-Mb/s的以太网端口可以以20-Mb/s的速率传送数据——其中每个方向各为10-Mb/s。

自然，我们必须保证相互作用设备的MAC层支持这种特殊的模式。当只有第一个节点支持全

⊖ 在路由器中，早就使用这样的技术了，它可以在一个由分层原则组织成的网络中减少地址表的大小。

双工模式时，另一个节点就记录冲突并挂起它的操作；但是第一个节点将继续传送数据，此时是不会有人接收的。为了使节点支持全双工模式，需要针对节点MAC层逻辑操作所做的修改是极少的。它只需要取消以太网中的记录和冲突处理。在令牌环和FDDI网络中，在任何时刻，只要端节点需要，网络适配器和交换机端口就必须发送它们的帧而不用等待令牌的到达。事实上，在全双工模式下操作时，MAC层的节点忽略了不同技术下的介质访问方法。

当早期快速以太网和千兆以太网时，全双工模式已经获得了全面的权力并且变成了网络节点操作中的标准模式。目前，网络适配器可以支持这两种类型的操作。当连接到集中器端口时使用CSMA/CD访问算法，当连接到交换机端口时便以全双工模式运行。

15.4.2　在全双工模式中拥塞控制的问题

仅仅放弃共享介质访问算法的支持而不修改协议会增加交换机的帧丢失率。这是因为端节点向网络上发送的对帧流的控制丢失了。在半双工模式下，典型的共享介质网络，帧流是由访问共享介质的方法调节的。在转换到全双工模式后，节点就可以在需要时给交换机发送帧；因此，在这个模式下运行时，网络交换机可能会被拥塞并且没有减慢帧的流动的方法。

通常，拥塞是不会由已经被阻塞的交换机生成的（即，处理器性能不足以满足帧流的处理）。拥塞的真正原因是某个输出端口的有限带宽，而这个带宽是由协议的参数所定义的。

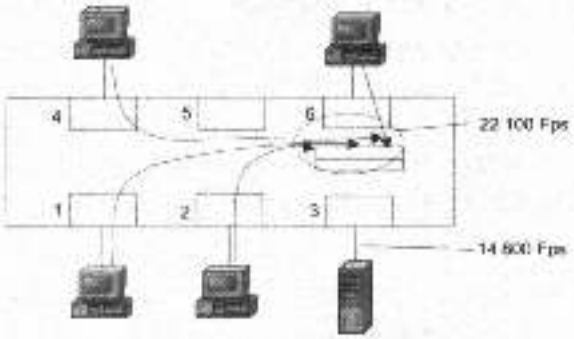


图15-20　不均衡的流量导致的端口缓冲溢出

因此，如果输出端口面的入流量是不均匀分布的，就很容易想像将会发生这样的情况：总强度超过了协议所定义的传送给交换机输出端口的最大流量。图15-20表示的正是这种情况。

这里，来自于端口1、2、4和6的64字节的帧流的总强度为20 100帧每秒，它被传送到端口3，端口3达到了150%的负载。自然地，当帧以20 100帧每秒的速率到达端口缓冲区，并以14 880帧每秒的速率离开这个端口时，输出端口的缓冲区将会被未处理的帧填满。

这不难计算：在这个例子中，一个100KB的缓冲区只要0.22秒就可以被填满（这样一个缓冲区最多可以存储1 600个64字节的帧）。即使把缓冲区大小增加到1MB，也只能将填满缓冲区的时间延长到2.2秒，这也是不能接受的。

这个问题可以用第7章讨论的拥塞控制方法来解决。

如你所知，有多种拥塞控制工具：交换机中的队列管理、反馈和带宽预留。在这些工具的基础上，可以建立一个支持不同类型流量的有效的系统。

在这一节，我们考虑1997年3月标准化为IEEE 802.3x规范的反馈机制。IEEE 802.3x的反馈机制标准仅仅用于双工模式下的交换机端口的操作。这个机制对于LAN交换机来说非常重要，因为它可以减少因为缓冲溢出而导致的帧丢失，无论网络是为各种流量保证不同的QoS支持，还是仅提供尽力服务。我们将在下一章详细讨论QoS的其他机制。

802.3x规范为以太网协议栈引入了一个新的子层：MAC控制子层。这个子层位于MAC层之上，并且是可选的（图15-21）。

这个子层的帧可以用于多种目的。然而，目前在以太网标准中只为它们定义了一个任务——即，在一个指定的时间开起所有其他节点的帧传输。

MAC控制帧与用户数据帧不同，因为它的类型/长度域总是包含88-08这个十六进制值。MAC控制帧是供全体应用使用的；因此，它的格式相当复杂（图15-22）。

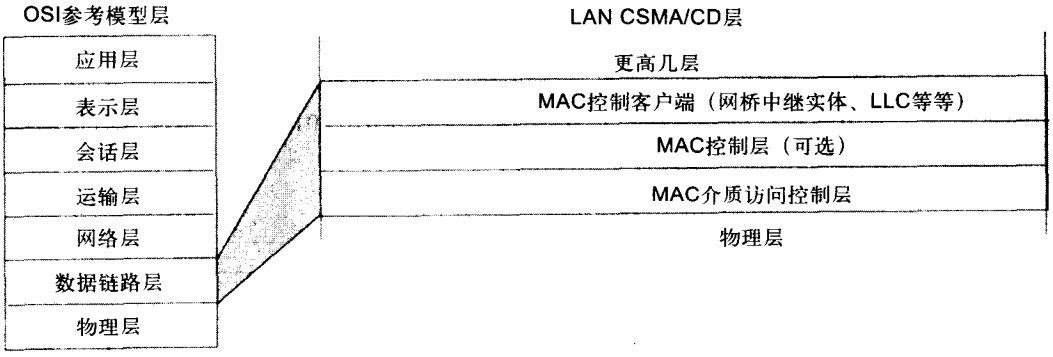


图15-21　MAC控制子层

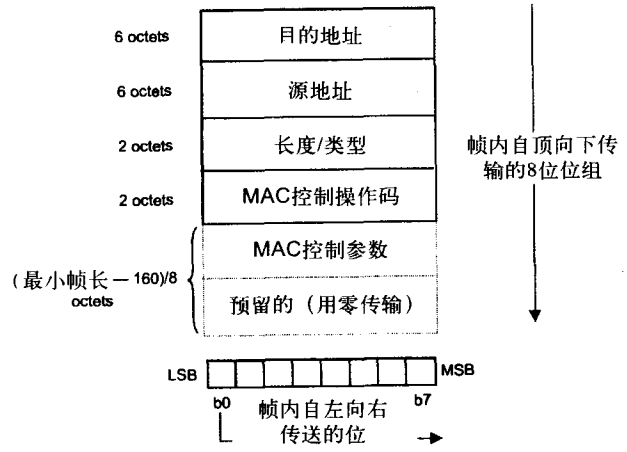


图15-22　MAC控制帧的格式

必要时，交换机使用MAC控制帧来临时挂起来自邻居节点的帧，从而降低它的内部队列的负载。

作为目的地址，可以使用为这个目的而预留的多播地址：01-80-C2-00-00-01。当邻居节点也是一个交换机时，这个方法是非常便利的（因为交换机端口没有唯一的MAC地址）。如果邻居是一个端节点，仍然可以使用这个唯一的MAC地址。

MAC控制操作码域指定了控制操作的编码。前面我们已经提到过，只定义了PAUSE操作，它的十六进制编码是00-01。

收到这个操作码的节点必须挂起与发送PAUSE帧的节点之间的帧传输，挂起的时间由MAC控制参数域指定。对于特定的以太网实现，这个时间单元是512比特间隔；可能的挂起间隔从0到65 535不等。

如上所述，这个反馈机制与第7章中的分类反馈类型2相似。当然，它也有自己的特点，在这种机制中，常使用两种操作：挂起帧传输和恢复帧传输。这是实现分组交换网络的最早的一个协议中的机制——这个协议也就是，X.25协议中叫做LAP-B的那个。

15.4.3　10G以太网

10G以太网标准只定义了操作的全双工模式；因此它只能用于交换LAN。

形式上，这个标准被叫做IEEE 802.3ae，它是802.3标准的一个补充。这个对于以太网家族的补充描述了七个新的物理层规范，它使用一种新的协调子层与MAC层进行交互（图15-23）。这个

子层给所有不同的10G以太网物理层提供了一个统一的接口，名为扩展吉比特介质独立接口（XGMII），它为4个字节的并行交换做好了准备。

从图15-23可以看出，10G以太网标准有3组物理接口：10GBase-X、10GBase-R和10GBase-W。它们的不同在于使用的数据编码方法：10Base-X方法使用8B/10B编码，其他两组使用的是64B/66B编码。三组都使用光纤介质进行数据传输。

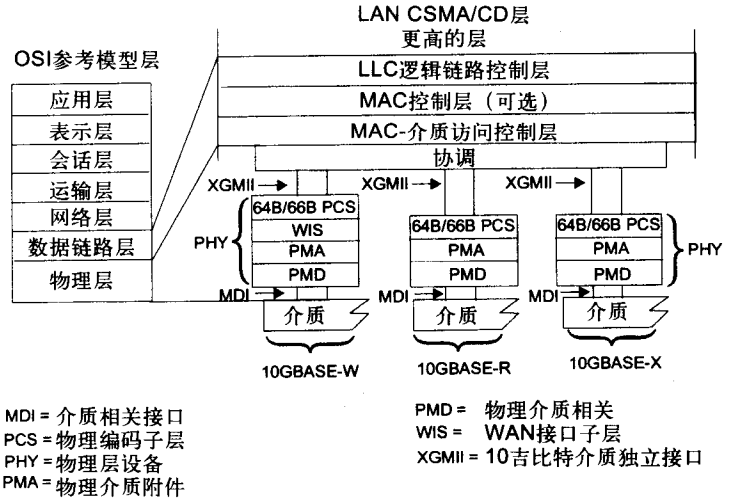


图15-23　10G物理接口的三个分组

10GBase-X包含一个物理介质相关（PMD）子层的接口：10GBase-LX4。字母L表示信息通过波长的第二个透明范围（即，1 310nm）传输。同时信息使用4个波在各个方向传送（在接口名字中是用数字4来表示的），它是基于波分复用（WDM）技术上（图15-24）的复用。XGMII的四个流都是以2.5Gb/s的速率在光纤中传送的。

根据10GBase-LX4标准，发送端和接收端的最大距离，对于多模光纤是200~300m（取决于光纤的带宽）；对于单模光纤，最大距离是10km。

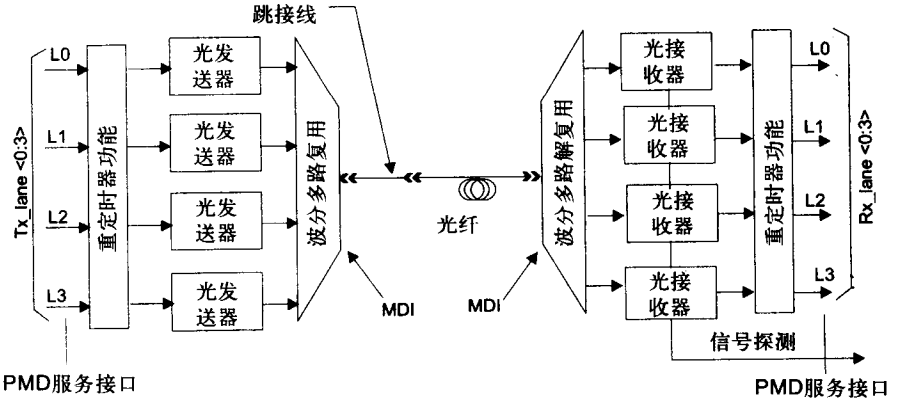


图15-24　使用WDM技术的10GBase-LX4接口

10Base-W和10Base-R组都由三个不同的RMD子层组成（S、L和E），取决于传输信息所使用的波长（分别是850nm、1 310nm和1 550nm）。因此，以下是它们的接口：10GBase-WS，10GBase-WL，10GBase-WE，10GBase-RS，10GBase-RL和10GBase-RE。它们都使用一个合适的

波段来传送信息。

10GBase-W组和10GBase-R组的区别在于：W组的物理接口保证了一个数据传输速率以及一个与Sonet STS-192/SDH STM-64接口兼容的数据格式。W组接口的带宽是9.95328Gps，数据传输有效速率是9.58464Gb/s（部分的带宽用于STS/STM帧头）。由于这一组接口的信息传送率低于10Gps，只能进行同类接口之间的相互操作，这意味着它不可能有与10GBase-RL和10GBase-WL这种的接口。

W组的接口与对应速率的SONET/SDH接口在电子特性上并不完全兼容。因此，为了保证使用SONET/SDH传输网络的10G以太网网络，传输网络的多路复用器必须装备有与10GBase-W规范兼容的特殊的10G接口。10GBase-W设备支持9.95328Gb/s的速率，这保证了使用SONET/SDH网络，以STS-192/STM-64帧的格式传送10G以太网流量的可能性。

运行在E透明窗口下的物理接口保证了数据传输距离可以达到40km。这不仅仅允许建立LAN甚至允许建立MAN，这在802.3标准的源文档的修正中被反映出来。

小结

- 在建立中型和大型网络时，网络的逻辑结构是必须的。只有对于由5到10台计算机组成的网络，使用公共共享介质才是可接受的。
- 将网络分成逻辑分段提升了网络的性能、可靠性、灵活性和可管理性。
- 为了网络的逻辑架构，我们使用了网桥以及它的继承者交换机。交换机仅使用了几个基于数据链路层协议的工具，就将网络分成多个逻辑分段。不仅如此，这些设备还不需要配置。
- 交换机用来建立地址表的被动方法就是追踪通过的流量。这使得它不能在包含闭合环的网络中运行。基于交换机网络的另一缺点是，无法保护网络不受广播风暴的袭击，这是因为它们的操作算法使得它们不得不传送这些广播信息。
- 交换机的使用使得网络适配器可以使用LAN协议（以太网、快速以太网、千兆以太网、令牌环和FDDI）的全双工操作模式。在这种模式下没有访问共享介质的步骤，并且数据传输速率加倍。
- 在全双工模式下，交换机的过载可以使用802.3x 标准中描述的反馈方法来预防。它临时挂起来自过载交换机的最近邻居的帧传输。
- 在交换机操作的半双工模式下，交换机使用两种方法来控制帧流量：主动的介质抢占和背压方法。这些方法的使用使得可以通过将多个传送帧和一个接收的帧交错来对流进行充分灵活的控制；
- 交换机性能的主要特征是帧的过滤率、帧转发率、以兆位每秒度量的所有端口的性能以及帧传输延迟；
- 交换机的性能取决于交换类型——直通型还是完全缓冲型——取决于地址表的大小和帧缓冲区的大小；
- 交换机可以使用多种标准过滤所传送的流量，这些标准既要考虑任意域的值，也要考虑源地址和目的地址。然而，在数据链路层，指定用户自定义过滤器的方法是相当复杂的。为了掌握这个方法，管理员必须对协议有充分的了解并且大量执行确定需要的特征值在帧的哪个位置的艰难任务。

复习题

1. 列出基于共享介质的网络的主要局限性。
2. 为什么在以太网中延迟急剧增长起始点的介质利用率要比FDDI网络和令牌环网络的低？
3. LAN交换机的优点有哪些？
4. 交换机的转发表是建立在下面哪个的基础之上？

a. 源地址

b. 目的地址

5. 可不可以这样说：当把共享介质分成两个分段时，每个分段的负载也减少到原来的一半？
6. 建立在运行透明网桥算法的交换机的基础上的网络中，闭合环路的存在会有哪些负面影响？
7. 限制转发表中条目的生存时间的目的是什么？
8. 比较透明网桥算法和SRB算法。
9. 在交换机中用户自定义的过滤器的目的是什么？
10. 当在交换机中建立用户自定义的过滤器时可以使用什么样的参数？
11. 转发速率可以超过过滤速率吗？
12. 什么是非阻塞交换机？
13. 一个非阻塞交换机会因为队列溢出而丢失分组吗？
14. 在内部队列溢出时，交换机会使用什么机制？
    - 流量整形
    - 基于PAUSE帧的反馈
    - 背压（人工的冲突）
    - 划分优先级
15. 10G以太网技术可以使用共享介质吗？
16. 物理接口的哪个特性对应于10GBase-LX4规范命名的数字4？
17. 可以直接将一个带有10GBase-WI接口的LAN交换机与SDH多路复用器的STM-64端口相连吗？
18. LAN技术的哪一个特征简化了以太网、令牌环和FDDI协议之间的转换？
19. 在什么情况下FDDI帧不能被转换成以太网帧？
20. 指出交换矩阵的主要类型。
21. 交换矩阵的主要缺点是什么？
22. 共享存储器与交换矩阵相比有哪些优势？
23. 为什直通交换在交换机的应用上具有局限性？
24. LAN地址的数目超出了交换机地址表的大小会怎么样？

练习题

1. 用户自定义的过滤器包括逻辑条件和如果条件满足则必须对帧执行的操作。用公式表示下列帧的过滤器条件：来自于计算机*A*的帧，*A*的MAC地址是06 DB 00 34 5E 27；来自于*B*的帧，*B*的MAC地址是CC 33 00 D5 43 4D；前往服务器S的帧，S的MAC地址是CC 33 00 65 44 AA。
2. 提高基于共享介质的网络的性能是有必要的（图15-25）。你只可以支配一个交换机，这个交换机有两个1 000Mb/s的端口和8个100Mb/s的端口。如果可以继续使用一开始就有的集线器，你怎么改造这个网络？

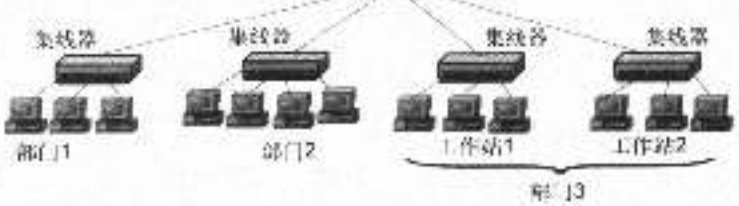


图15-25　需要升级的网络

第16章 交换LAN的高级特性

16.1 引言

为了建立更大规模网络，交换机可以将共享介质分割成多个部分，或者抛弃介质共享原则而使用交换LAN。然而，随着网络规模的增长，产生了其他的问题，而这些问题是上一章讨论的仅基于透明网桥算法的交换机所不能解决的。首先，可靠性的问题仍然没有得到解决，交换LAN的树状拓扑极易受到攻击。例如，任何交换机或者通信链路的失效将导致连通性的丢失。如果网段交换机失效，网络就会被分成两个或多个网段。

树状拓扑的局限性可以通过使用附加的交换机制来避开，这种机制为LAN提供了一些先进特性。例如，生成树算法（STA）在交换LAN中被广泛使用。一旦出现交换机或者通信链路的失效，这个算法就会自动找到一个新的树状拓扑变体，这样保证了网络的容错性。STA与透明网桥算法在同一时期被开发出来（也就是，在20世纪80年代早期）；并且从那时开始，它就被成功地用于LAN。

另一个机制是使用可替换路由，这个机制开发时间相对较晚，与开始大规模使用交换LAN的过程同步。这种链路聚合机制允许多个物理链路组合成一个逻辑信道。这提升了网络的性能和可靠性。

LAN交换机的新的先进功能可以使得许多服务质量（QoS）的通用机制可以实现，这主要是用于不同种类的流量：优先级和加权队列、反馈和资源预留。

尽管由于使用了由STA和链路聚合机制增加的新特性，取得了一点进展，但是，没有使用路由器而仅建立在交换机基础上的LAN，它的特征是有很多的局限并具有一定的问题。使用交换LAN的一个重要的高级特性——虚拟LAN（VLAN）技术，可以部分地解决这些问题，这项技术大大地简化了网络中用路由器的使用。VLAN技术允许将LAN分割成一些独立的逻辑网段。这是使用交换机配置实现的（例如，仅仅是编程实现，而不是物理地连接或断开电缆插头）。这些独立的网段可以使用网络层协议连接到互联网中。以编程的方式将网络分割成多个网段，允许通过将计算机从一个网段移到另一个网段，来简单快捷地改变网段的组成。

16.2 生成树算法

在这些LAN中，无论是技术还是设备都仅仅实现了ISO/OSI模型的第一和第二层的功能，使用可替换路由的问题有它自己的特征：基础协议仅支持树状拓扑（也就是，不包含任何环路的拓扑）。

当所有可选的链路不适合于树状拓扑的框架时，LAN使用**生成树算法**（spanning tree algorithm，STA）自动转换到一个保留的状态中。实现这种算法的协议称为**生成树协议**（spanning tree protocol，STP）。

早在1983年STA就被开发出来了。IEEE采纳了这个算法并且将它收录在描述透明网桥算法的802.1D规范中。虽然这个算法最初是用于网桥的，且网桥已被认为是“恐龙”级别的通信设备，但是这个算法仍广泛用于当前LAN中最普遍的通信设备——交换机中。STA使得网络设计者可以不使用路由器而仅在基于交换机的基础上建立大规模的LAN。因为预留LAN的使用，这种LAN具有高度可靠性。

通常，网络设备的制造商在用于这种网段的交换机中实现了STA算法，这些网段的特征是对可靠性的高要求，这些交换机有主干交换机和应用于部门以及大型工作组的交换机。

16.2.1　必要的定义

在图16-1中，STA用图的形式描绘了由交换机和网段组成的网络。节点是交换机和网段。

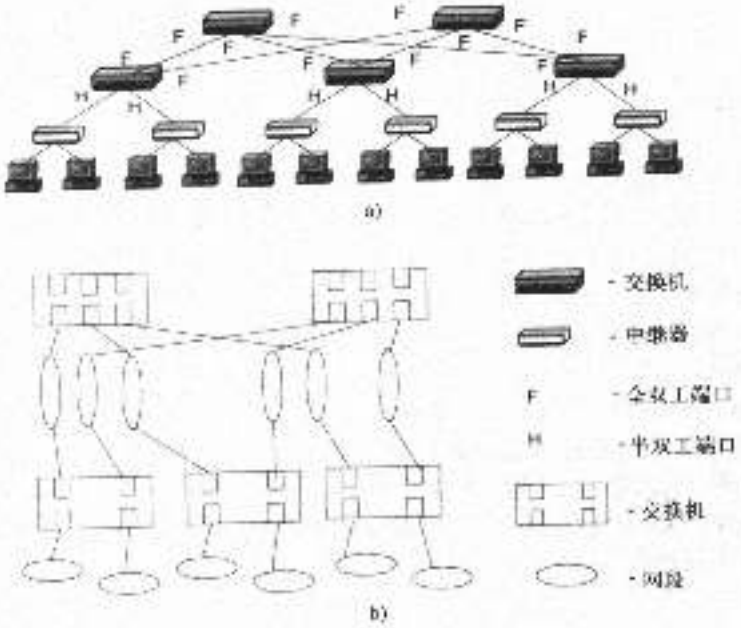


图16-1　基于STA的网络的形式化表示

网段（segment）是网络中一个不包含交换机和路由器的连通部分。它可以用于共享（当STA被研发出来的时候，它成为了网段唯一的类型），也可以包含物理层的设备例如中继器或集线器，这些设备对于交换机来说都是透明的。现在，网段通常是两个交换机的相邻端口间的一个双向的点到点链路。

STA保证建立链路的一种树状拓扑，从每个交换机和网段到某个专门的**根交换机**（root switch），也是树的根，树状拓扑有且只有一条最短路径。路径的唯一性保证了链路中没有环路；距离的最小性允许为流量建立一条从网络外围通向网络主干的路径，在这里网络主干是根交换机。

为了测量距离，STA使用路由协议常用的**度量**（metric）作为距离的计量单位，即，与网段的带宽成反比的值。在STA中，度量也被定义为网段的指定代价。它的值被计为传输1比特信息所需的时间，且用10nesc作为计量单位。因此，对于10Mb/s以太网网段，指定代价为10；对于100Mb/s以太网网段，指定代价为1；而对于令牌环16Mb/s网段，这个值是6.25。考虑到网络速度在不断地增长，出现了一种重新修订的单位比例：10Mb/s~100，100Mb/s~19，1Gb/s~4以及10Gb/s~2。

交换机标识符（switch identifier）是一个8字节的数字，它的后六个字节包含了实现STA的交换机控制单元的MAC地址。前面我们曾提到，交换机和网桥的端口不需要使用MAC地址来实现它们的主要功能，所以这并不是一个端口的MAC地址。交换机标识符的前面两个字节是手工配置的，正如你后面将要看到的，它使得网络管理员可以控制选择根路由器的进程。

交换机的**根端口**（root port）就是交换机中离根交换机具有最短路径的端口（更准确地说，是与根交换机的任何端口具有最短路径的端口）。

端口标识符（port identifier）是一个2字节数字。后一个字节表示的是交换机中该端口的序号，前一个字节是由网络管理员手工设置的。

指定端口（designated port）是本网段中的所有交换机的所有端口中离根交换机距离最近的端口。

网段的**指定交换机**（designated switch）就是网段中具有指定端口的交换机。

网桥协议数据单元（Bridge protocol data unit，BPDU）是一种特殊的分组，交换机之间通过定期交换这种分组来实现对树结构的自动探测。BPDU携带有关交换机和端口标识符的数据以及从根交换机的路径代价的信息。在STA中，BPDU分组生成的间隔称为*hello*间隔。这个间隔由网络管理员设置，通常是1到4秒。

16.2.2 构建生成树的三步过程

图16-2给了我们一个用于说明建立生成树过程的示例网络。

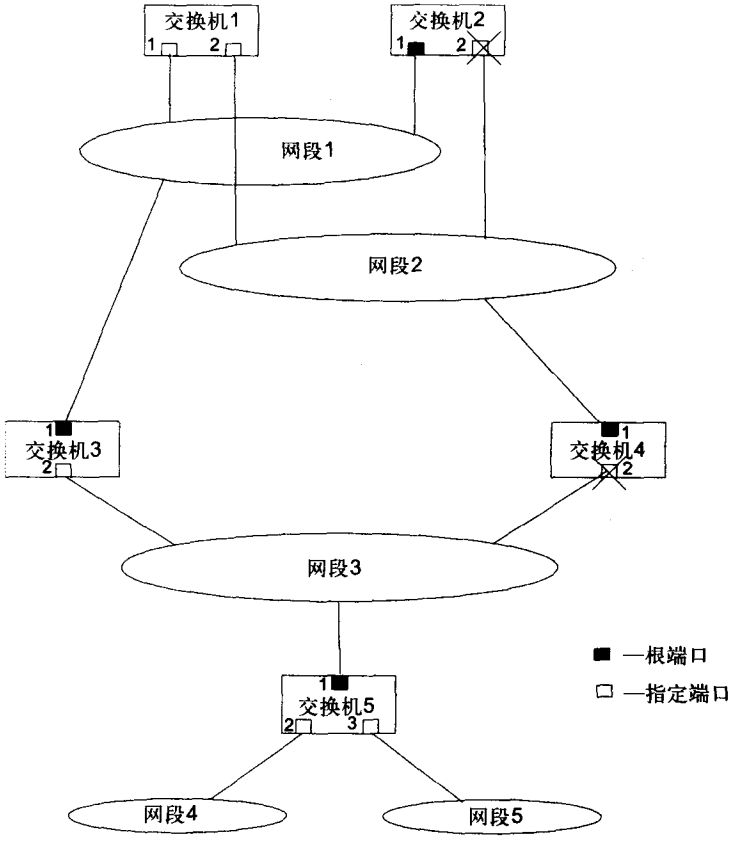


图16-2 根据STA建立生成树的示例

STA通过三个步骤来确定网络的动态结构。

步骤一。为要建的树选择根交换机。根据STA，这个根交换机由具有最小标识符值的交换机担当。如果网络管理员不参与这个过程，那么将会随机地选择根交换机。事实上，具有MAC地址值最小的控制单元被用于担当这个任务。自然，这种选择远远不是最佳选择。例如，如果交换机5被选为根交换机（图16-2），那么，大部分的流量就要经过许多中间网段和交换机来传递。因此，网络管理员是不应该让这个过程正常进行的。如果网络管理员可以改变这个过程并且基于合理的考虑选择根交换机，那么情况会变得好很多。这是通过对交换机标识符中最重要的几个字节进行适当的设置来实现的。以这种方法进行处理，就能选择出在段间连接中，占据中心位置的交换机。假设图中的数字对应于交换机的标识符。此时，交换机1将被选做根交换机。

步骤二。每个交换机都要选择根端口。与根交换机的距离是根据从根交换机发来的BPDU确定

的。基于这些分组的数据，每一个交换机都可以确定它所有的端口中到根交换机的最小距离。每一个交换机分析接收到的BPDU数据分组，将BPDU数据分组中指定的根交换机的路径代价加上上次接收分组的网段的指定代价，然后再转发这些BPDU分组。这样，数据分组从根交换机到达不同交换机时，BPDU中的距离值就会不断地增加。例如，假设例子中的所有网段是10Mb/s以太网的网段，交换机2收到一个来自于网段1的BPDU分组，分组中的距离原来的设置为0，交换机2就要将这个距离值增加10。

当转发分组时，每一个交换机“记住”其每一个端口在所有收到的BPDU数据分组中与这个根的距离的最小值。完成了定义生成树结构的过程以后，每一个交换机也找到了它的根端口（也就是，离根具有最短距离的端口）。

如果度量值相等，那么交换机和端口标识符就用于处理这种二义性。优先选择具有最小标识符的端口和交换机。例如，对于网段3，有两条路径通往根交换机1，这两条路径的度量值是相同的，第一条路径是通过交换机3，第二条路径是通过交换机4。那么选择通过具有较小标识符的交换机（即，通过交换机3）的路径。在这个例子中，交换机的端口号也是相等的，然而，在进行比较时，我们总是先考虑交换机的标识符，其次考虑端口号。

在图16-2的例子中，交换机3选择端口1作为它的根端口，因为这个端口与根的最短距离是10（通过网段1接收的来自根交换机的具有这样一个距离的BPDU分组）。交换机3的端口2，基于接收的分组，发现与根的最短距离是20。这段路程是由根交换机的端口2出发经由网段2，然后通过交换机4和网段3。当交换机2选择根端口时，就遇到了端口1和端口2与根的距离都是10的情况。端口1经由网段1接收来自根路由器端口1的分组；端口2经由网段2接收来自根路由器端口2的数据分组。由于端口1的标识符小于端口2，所以端口1被选作为根端口。

步骤三：指定端口选自于一个网段之中的所有交换机的所有端口，具有指定端口的交换机成为那个网段的指定交换机。选择指定端口与选择根端口相似，使用了分布式过程。网段中的每一个交换机首先将根端口排除在考虑之内，因为对于连接到根端口的网段，往往会有另一个交换机与根更近。对于所有剩下的端口，它们与根交换机的最短距离（在加上这个网段的通话时间之前）首先与这个交换机的根端口与根交换机的距离做比较。如果这个端口所有方向的距离都比交换机根端口的距离要大，那么对于连接到这个端口的网段，到达根端口的最近距离就需要经过这个交换机。因此，这个交换机就成为指定交换机。这个交换机让它的所有满足这个条件的端口成为指定端口。当有多个端口与根交换机都有相等的最小值时，就选择具有最小端口标识符的端口。

在这个例子中，交换机2通过端口2检测到一个具有最小距离0的分组（这些分组来自于根交换机1的端口2）。由于交换机2的根端口到根的距离是10，所以这个交换机的端口2就不能成为这个网段的指定端口。

默认情况下，交换机有15秒的时间来完成这全部的3个步骤。我们可以假设在这期间，每一个交换机接收了足够多的BPDU，使得它可以确定所有端口的状态。

所有其他的端口，除了根端口和指定端口以外，都被转换到阻塞状态（在这个插图中，这些端口都被画了叉），这样建立生成树的过程就完成了。可以用数学证明，这种选择活动端口的方法消除了网络中的环，并且剩余的连接形成了生成树（假设它可以用已有的网络链路来创建）。

说明　一般情况下，根据STA选择的树状拓扑在所有可能的流量传输路径中并不是最优的。例如，在上面的例子中，当从网段3向网段2传送分组时，流量沿着以下路径传输：交换机3—网段1—交换机1—网段2。这条路径的度量值是30。如果交换机4的端口2没有被阻塞，就会有一条更短的路径：通过交换机4的路径。这条路径的度量是20，这比前一条路径要好。这条路径是有可能存在的，如果网段2选择让到根交换机的最短路径经过交

换机4而不是交换机3。这可以通过给交换机标识符的前两个字节设置一个适当的值来实现。然而，如果选择了这条路径，那么从网段4到网段1的路径就不再是最优的了。

建立了生成树后，交换机开始接收数据分组（不转发），并且在它们的源地址的基础上建立转发表。这是一种普通的、透明网桥-学习模式，它不能提前被激活，因为端口事先不能确定自己是否仍然是根端口或成为一个将传送数据分组的指定端口。默认情况下，这个“学习”过程也将持续15秒。同时，端口继续参与STA操作，这也表示到达的BPDU如果具有更好的参数，则自动地把这个端口转换成“阻塞”状态。

只有在度过一个两倍于预定义的超时值的时间间隔之后，端口才转换为“转发”状态，并且开始根据生成的转发表来处理分组。这里要注意的是，这个转发表为了显示网络结构的变化，会继续被修改。

在正常运行期间，交换机继续每隔一个hello间隔就生成配置BPDU分组；其他交换机通过它们的根端口接收这些分组，然后通过指定端口转发这些分组。交换机可能会缺少指定端口（如，例子中的交换机2和交换机4）；然而，因为它们的根端口持续接收BPDU分组，所以它们也参与了STA协议的操作。

如果任何一个网络交换机的根端口在消息的最大**生存时间**（time to live，TTL）之后（默认情况下是20秒），仍然没有接收到BPDU分组，那么它就启动一个新的建立生成树的程序。在这种情况下，交换机生成一个BPDU分组，在这个分组中，它指定自己为根并且将这个信息告诉给所有的交换机。所有其他的消息TTL计时器到时的网络交换机以类似的方法处理。结果，就选择了一个新的活动结构。

16.2.3 STA的优点和不足

与大部分简化的算法相比，STA的一个主要优点就是，当它的邻居设备失效时，就会有一个唯一的预留连接来接管，这个连接重组网络的结构，不仅会考虑到它的最近邻居的链路，还会考虑远程网段的链路状态。

这个算法的一个缺点是，当网络中具有大量的交换机时，确定一个新的动态结构所需的时间太长了。如果网络使用默认的超时值，那么转化成一个新的结构需要的时间可能会超过50秒。其中20秒用于确认与根交换机的连通性（在标准的STA中，计时器超时是得到这些信息的唯一方法），另外还需要两个15秒用于进入“转发”状态。

目前非标准的STA算法相当多，它们允许通过增加算法复杂度来减少重构的时间。这可以通过增加新型的控制信息来实现。在2001年，开发出了生成树的一种新版本：IEEE 802.1w规范，仍是为了加速协议的运行，但是是以标准的方法提高协议的操作速度。

16.3 LAN中的链路聚合

16.3.1 干线与逻辑信道

把两台通信设备之间**链路聚合**（Link aggregation）成一个逻辑信道，是使用LAN中冗余可替换链路的另一种形式。

链路聚合技术和前面讨论的STA技术有原则上的不同。

- STA将冗余链路都作为一种热保留，在保证网段连通性的前提下，只让最少的信道处于活动状态。这样，网络可靠性增加了，但是它的性能还是不变。
- 当使用链路聚合时，所有的冗余链路也是活动的，这样既增加了网络的可靠性也提高了网络的性能。

如果这种聚合信道或者称做干线中的某个部件失效，那么流量将分配给剩余的链路中（图16-3）。

在这个例子中，出现这种情况的是干线2，它的一条链路（中间的）失效了，所以所有的帧都是由剩下的两条链路传输的。这个例子表现了可靠性的增加。

现在考虑链路聚合是如何提升网络的性能的。例如，在图16.3中，交换机1和交换机3由3个并行的链路连接，与不允许并行链路的标准树状拓扑的标准变体相比，链路聚合增加了这个网段3倍的性能。在某些情况下，通过链路聚合实现的交换机间连接性能的提升，要比仅仅将物理链路替换成一个更快的物理链路，其效率要高得多。例如，尽管以太网家族给我们提供了很多种物理链路的速度——从10Mb/s到10Gb/s不等，但是通过转换为一个更快的以太网标准来实现速度10倍的提升并不一定需要，而且也不合算。例如，如果安装在网络中的网络交换机没有提供增加安装千兆以太网端口模块的可能性，那么将一些链路升级到1 000Mb/s，就需要彻底地更换交换机。另一方面，已经存在的交换机可能还有空闲的快速以太网端口。因此，可以通过将6个快速以太网连接聚合起来把数据速率提高到600Mb/s。

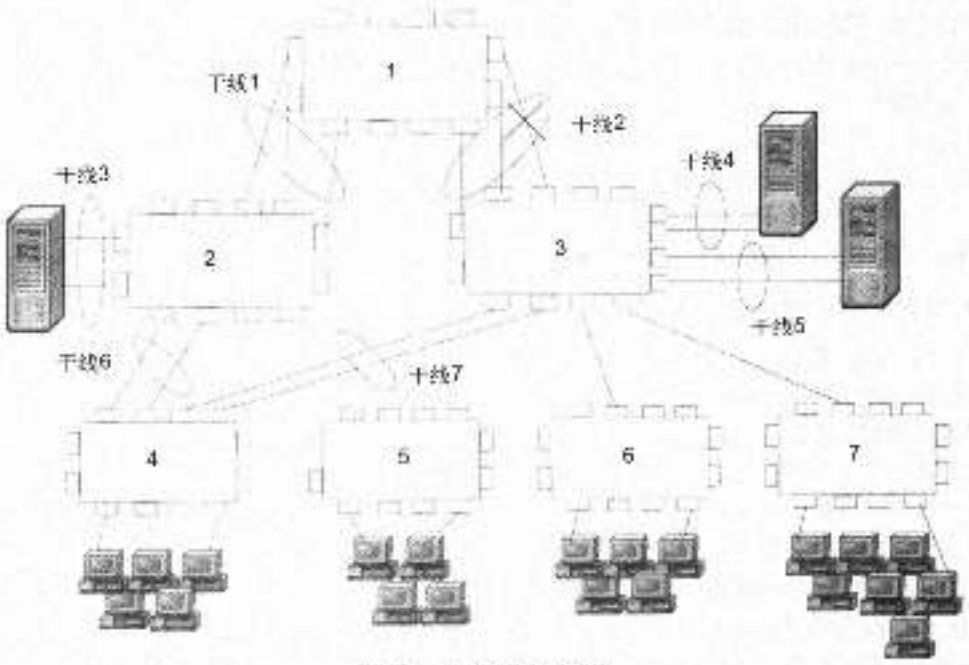


图16.3　物理链路的聚合

链路聚合是由第6章中描述的使用可替换路由的第三种方法（"网络事先找到了两个路由，但是只使用其中一个"）[⊖]的一般化形式。在这种情况下，不是两个路由，而是找到了N（N大于等于2）个路由，但是每一个数据流只能使用其中一个。当这个路由失效时，由于这个故障所影响的数据流就转换到剩余的$N-1$个可用路由中的任意一个。

链路聚合用于两个交换机端口之间的链路，用于交换机和计算机之间的链路。更通常的是，这种变体用于快速或者企业关键服务器中。此时，组成某个干线的所有网络适配器或者交换机端口共享同一个网络地址。因此，干线的端口对于IP或任何其他网络层协议来说是不能区分的，这与作为链路聚合基础的一个"统一的逻辑信道"的观念是一致的。

当今使用的链路聚合方法几乎都有一个严重的局限性：它们只考虑两个邻居网络交换机之间的链路，而忽略了任何发生在网段之外的链路。例如，干线1的操作与干线2的操作并不协调。因此没有考虑到交换机2和交换机3之间存在着一条备通链路，而正是这条链路创建了一个包含干线1和干

⊖　请参阅6.4.3节。

线2的环路。正因为如此，如果网络管理员想使用连接网络节点的所有拓扑能力，那么链路聚合必须和STA一起使用。对STA而言，干线必须看起来像一条链路，那么，STA的操作逻辑就仍然有用。

目前有许多链路聚合机制的专有实现。无疑，最流行的实现的属于LAN设备工业中的领袖。这些流行的实现包括Cisco的快速以太信道和吉比特以太信道、Nortel的多链路干线以及Intel的可适应负载平衡。IEEE 802.3ad（链路聚合）总结并推广了这些方法。

16.3.2 消除帧的生育

现在，让我们考虑当交换机的端口组成干线时，其操作的细节特性。在图16-4显示的一小段网络中，两个交换机，交换机1和交换机2由4个物理链路相连。干线既可以是单向的也可以是双向的。每一个交换机都仅仅控制帧发送，并决定传送的输出端口。因此，如果两个交换机都认为连接它们的链路是干线，那么这个干线就是双向的，否则就是单向的。

图16-4列出了交换机1与并行链路相关的操作。如果交换机不认为这些链路是聚合链路，以下两种帧就会出现问题：

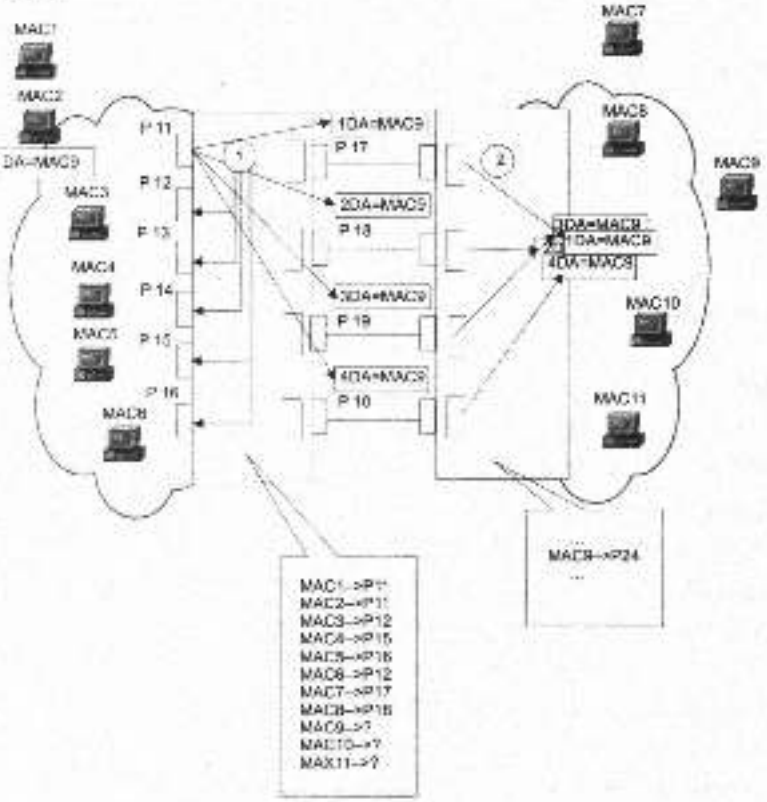


图16-4 当使用交换机间的并行链路时，具有未知地址的分组的生育现象

- 携带交换机还没有学习到的特殊地址的帧
- 携带广播或者多播地址的帧

透明网桥算法要求交换机将具有未知地址（也就是，在转发表中不存在的）的帧传送给除了接收该帧的端口以外的所有端口。如果具有并行链路，那么这种帧就会大量产生。它的拷贝的数目与并行链路的数目相同。在前面提供的例子中，交换机2就会收到原始帧的4个拷贝。

同时，帧会因为在交换机之间进行不断的循环而陷入无休止的回路。注意，这种帧是不能从网

络中被移除的，因为数据链路层协议往往缺少*TTL*域，虽然这个域常用于IP或者IPX这样的上层协议中。

在任何情况下，具有未知地址的帧由于增加了帧的数目，从而加重了网络负载。这可能会导致拥塞、延迟以及数据丢失。除了网络负载的增加外，帧的副本还会导致许多上层协议的低效操作。例如，使用TCP的网络节点，它使用ACK副本作为网络拥塞的间接信号。

具有广播地址的帧会导致更多的问题，因为它们必须传送给除了输入端口以外的所有其他端口。因而，网络会被许多不需要的流量所淹没，这种负载相当沉重，帧将会陷入永无休止的循环。

如果帧中具有已知且唯一的地址，就不会产生这种问题。这是因为交换机只需把帧传送到一个唯一端口即可，也就是交换机收到源地址为该地址的帧的端口。

聚合机制的开发者已经开始研究当帧中具有未知地址、广播地址或者多播地址时所导致的问题。解决这个问题的方法很简单。在转发表中，并行链路连接的所有端口被指定为一个*逻辑端口*，而不是一些*物理端口*。在图16-4中，转发表包含了一个逻辑端口，AL11，而不是P17、P18、P19和P10端口。所有通过交换机2的节点，它们的地址都被映射到这个端口。同时，学习到一个帧携带的新地址后，该帧可以来自于干线内任何一个物理端口，都将在转发表插入一个新条目。新条目包含了逻辑端口的标识符。那些目的地址被学习并被映射到逻辑端口标识符的到达帧，仅被传送给交换机的一个包含在干线中的输出端口。交换机使用相同方法来处理未知的、广播或者多播地址——即，它只为帧传输使用一条链路。帧处理逻辑的这个改变不适用于不在干线中的交换机端口。例如，交换机1总是传送具有广播或者未知地址的帧给端口P11~P16。

正是由于这个决策，帧就不用进行不必要的复制，前面讨论的问题也就不会出现了。

说明　事实上，这只有在两个交换机都将并行链路看成干线的情况下成立。因此，为了充分使用干线的特性，必须同时配置其两端的交换机。

16.3.3　端口选择

仍然存在一个问题：通过干线进行帧转发时需要使用哪一个交换机端口？

这里有多种答案。提高两个交换机之间、或者一个交换机和一个服务器之间的网段的整体性能是链路聚合的目标之一。因此，需要实现在干线端口之间的**动态帧分配**（dynamic frame distribution），需要考虑到每个端口的当前负载。这表示新到达的帧将被传送给具有最低负载的端口（例如，具有最短队列的端口）。看起来动态的帧分配方法一定会实现整个干线的最大性能，因为它考虑了每个端口的当前负载并且保证干线中所有链路上的负载平衡。

然而，这个论述并不是永远正确的，因为它没有考虑上层协议的行为。如果两端节点的会话分组的到达顺序与它们的发送的顺序不同，那么某些协议的性能可能会被大幅消减。如果同一个会话的两个或两个以上的顺序帧通过干线的不同端口传送就可能出现这种情况，因为这些端口的缓冲区中的队列可能具有不同的长度。因此，帧传输延迟也可能不同，所以后发送的帧可能会先于先发送的帧到达目的节点。

因此，大多数链路聚合机制在端口间使用**静态帧分配**（static frame distribute）而不是动态分配方法。静态帧分配采取在干线中专门为两个节点间建立会话的帧流分配一个端口。该会话的所有帧通过相同的队列，这就保证了它们到达目的节点的顺序与它们的发送顺序相同。

通常，当使用静态分配的时候，为某个会话选择的端口是基于到达的分组的某些属性。最常见的是，将目标MAC地址或者源MAC地址，或者两者一起，用作这种属性。在流行的Cisco快速以太信道交换机的实现中，这种交换机属于Catalyst 5000/6000家族，对目的和源MAC地址的最后两位有一个专门的OR（XOR）操作，这个操作用于选择干线的端口号。操作的结果可能会产生下

列某个值：00、01、10或者11。这些值都是常规的端口号。图16-5展现了使用快速以太网信道机制的一个例子。在这个例子中，两端节点间的会话的数据流的分发是随机的。由于这种分发并不考虑每一个会话的实际负载，所以干线的总带宽的使用可能是低效的，特别是在会话的强度差别很大的情况下。此外，这个算法甚至并不考虑各个端口上的会话分布的均衡。网络中的MAC地址的随机设置可能会导致某个端口传送几十个会话，而另一个端口仅传送两、三个会话。当使用这种算法的时候，只有在端口间建立许多网络计算机和会话时，才能实现各个端口间的负载均衡。

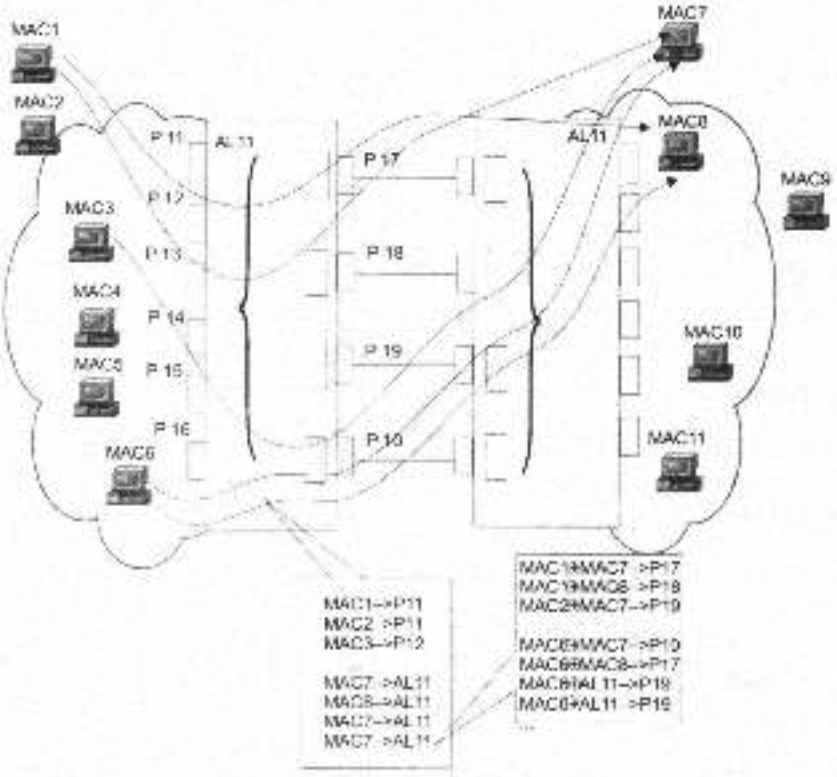


图16-5 使用快速以太网信道机制的网络

还有其他一些在端口间分配会话的方法。例如：可以根据封装在数据链路层的帧中的分组的IP地址，或者应用层的协议类型来执行这项任务。例如，使用一个端口传输电子邮件，使用另一个端口传送Web流量等等。这个例子是很有用的，那就是：将会话分配给学习到它的MAC地址的那个端口。在这个例子中，会话的流量在两个方向上都通过相同的端口。

在IEEE 802.3ad规范中描述的创建聚合信道的标准方法，假设可以基于分布在多个交换机上的物理端口创建一个逻辑端口。为了让交换机可以自动知道某个物理端口属于哪个逻辑端口，这个规范提供了一个特殊的服务协议：**链路控制聚合协议**（link control aggregation protocol，LCAP）。从而，就可以组织聚合链路的结构，这些结构不仅提升了两个交换机之

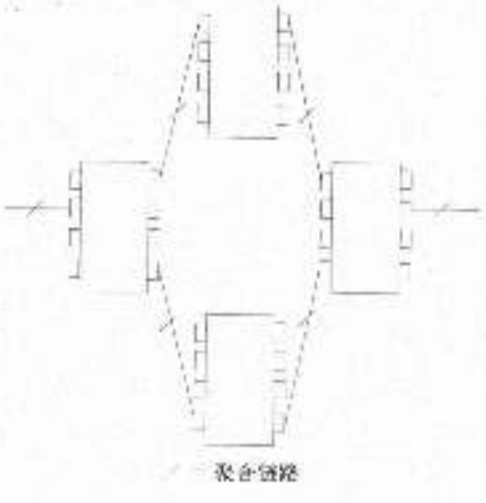


图16-6 分布式链路聚合

间的网段的网络容错能力，还可以应用于更复杂的拓扑（图16-6）。

当干线中的一条聚合链路失效时，分派给这个会话的对应端口的所有分组都会被转送给某个余下的端口。通常，恢复连通性的程序需要几毫秒到几十毫秒不等。因为在许多干线的实现中，映射到失效的物理链路上的MAC地址将被强制标记为未知。然后，交换机重复学习这些地址的过程。接着，地址再次被标记成已知，并且再重复一遍会话分配的过程。这次，只考虑可用的端口。由于在LAN的会话层协议中，超时设定很少有很长的，所以用于恢复损失的连接的时间也不能太久。

16.4　虚拟LAN

LAN交换机的一个重要的特征就是具有控制网段间的帧传输的能力。由于多种原因，在传送帧到指定的目的地址时，并不一定需要遵守访问权限或执行安全策略。

在第15章我们曾提到过，这些工作可以由用户自定义的过滤器来执行。然而，用户自定义的过滤器只能阻止到某个特殊的目的地址的帧传输。相反，广播流量可以传送到所有的网段。这是在交换机上实现的桥算法所要求的。因此，建立在网桥和交换机基础上的网络有时会被称做无力的，因为它无法阻止广播流量的传播。

VLAN技术允许管理者克服这个局限。

虚拟LAN（Virtual LAN，VLAN）是这样一些网络节点，它们的流量，包括广播流量，是与其他网络节点相隔离的。

这就表示在不同的虚拟网络间，基于数据链路层地址的帧传输是不可行的，无论这是哪一种地址——单一的、多播的甚至是广播的。同时，在VLAN中的帧根据交换技术进行传送（例如，仅仅传送给与帧的目的地址相关的端口）。

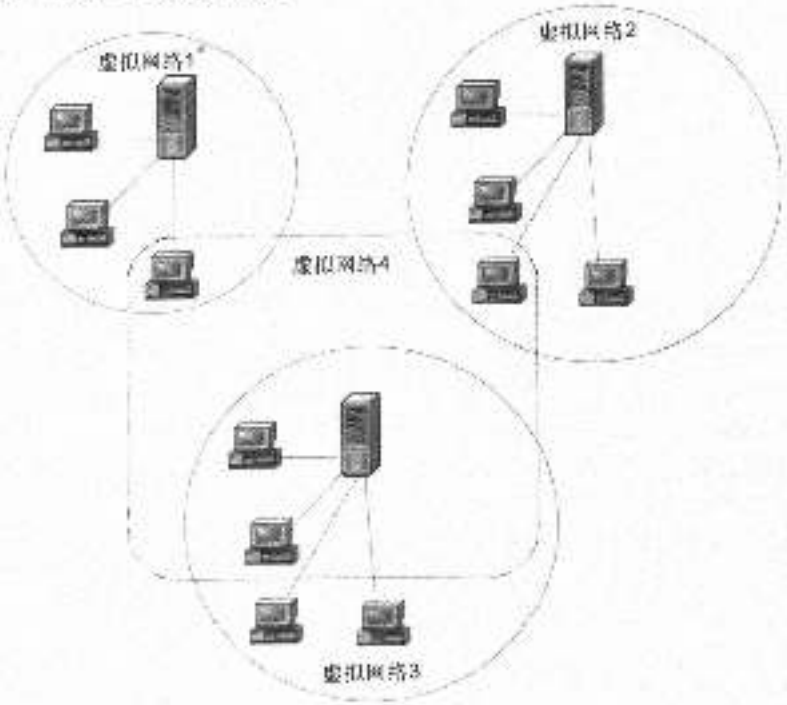


图16-7　VLAN

VLAN允许多重，也就是说一台计算机可以同时在不同的VLAN中。图16-7的例子中，电子邮

件服务器同时参与了VLAN 3和VLAN 4。这表示它的帧被交换机传送到所有参与到那些网络中的计算机。如果某个交换机仅仅在VLAN 3中，它的帧就不会到达网络4。然而，它可以通过公共邮件服务器与网络4中的计算机进行通信。这种方法并不能保证对于VLAN有一个完全的保护，因为例如发生在邮件服务器的广播风暴，也将会淹没网络3和网络4。

因此，虚拟网络组成了一个广播域，这个名字是由以太网中继器建立的冲突域类推得到的。

16.4.1 VLAN目的

VLAN技术的主要目的是建立独立的网络，这些独立的网络必须由实现某个网络层协议（如，IP）的路由器相连。这样的网络结构创造了更可靠的屏障，阻止了不需要的流量在网络间的传播。目前，我们一致认为每个大型网络都需要路由器，否则，像广播帧一样对交换机透明的错误帧流将会淹没整个网络，从而妨碍网络的运行。

由于交换机允许管理员不使用物理交换而仅通过变换配置来建立独立的网段，VLAN技术为建立一个内含许多子网的大型局域网提供了一个灵活的基础。

在VLAN技术出现以前，建立独立的网络要么使用物理独立的同轴电缆网段，要么使用建立在中继器或者网桥之上的彼此不相连的网段。后来，这些独立的网络通过路由相连组成了一个完整的互联网（图16-8）。

当使用这种方法时，网段结构中任何变化的引入，例如将用户移至另一个网络或者分割大型网段时，就需要对中继器的前面面板上的或者交叉面板中的连接器进行物理转换。这对于大型网络来说是很不方便的，因为这需要大量的手工操作，并且也很易出错。

多段集中器在一定程度上解决了这个问题（第14章曾提到过），它消除了网络节点的物理转换的需要。这样提供了在没有物理重接的条件下，对共享网段进行设计的可能性。

然而，使用集中器来解决改变网络结构的问题其实是很片面的，因为它对网络结构有极大的限制。由于这种集中器支持的网段的数目通常都不大，因此，希望使用这种集中器来给每个节点分配一个独立的网段是不切实际的（与可以这样做的交换机相比）。此外，当使用这种方法时，在网段间传输数据的工作虽然交给了高性能的路由器、交换机，但它们仍然"无法担当重任"。因此，建立在中继器基础上的具有结构交换的网络仍然依赖于许多节点间的介质共享机制，自然，它们要比建立在交换机基础上的网络的性能低很多。

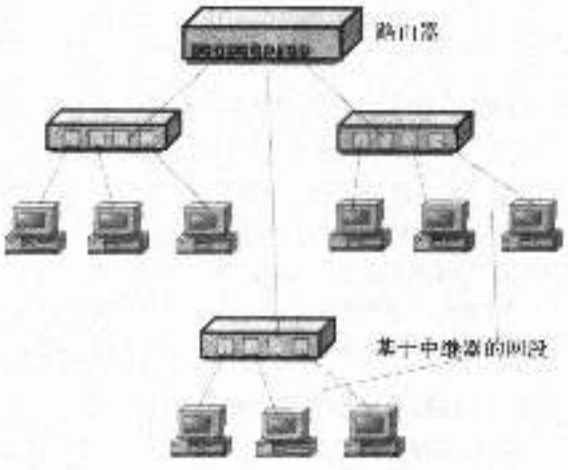


图16-8 基于中继器的多个网络组成的互联网

为了将多个VLAN连为一个互联网，就需要使用网络层工具，网络层的功能可以在一个单独的路由器上实现，或者也可以作为交换机软件的一部分。注意，此时的交换机已经变成了一个组合的设备，**第3层交换机**（layer 3 switch）。在本书的第20章中，我们将重点讨论第3层交换机。

构建基于交换机的VLAN和它们的操作的技术曾在很长一段时间里都没有被标准化，虽然它在不同的制造商的多种交换机中都有实现。直到1998年，IEEE 802.1Q标准被采用后，这一情况才得以改变，这个标准定义了构建与交换机支持的数据链路协议无关的VLAN的基本规则。

16.4.2　构建基于一个交换机的VLAN

当基于某个交换机构建VLAN时，需实现一个特殊的交换机端口聚合机制（图16-9）。此时，每个端口都赋给一个特定的VLAN。从属于另一个VLAN（如，VLAN 1）的端口传来的帧，永远不会被传送给不属于这个虚拟网络的端口。一个端口可以被赋给多个VLAN，然而，实际上这种情况很少发生。因为它消除了网络完全隔离的效果。

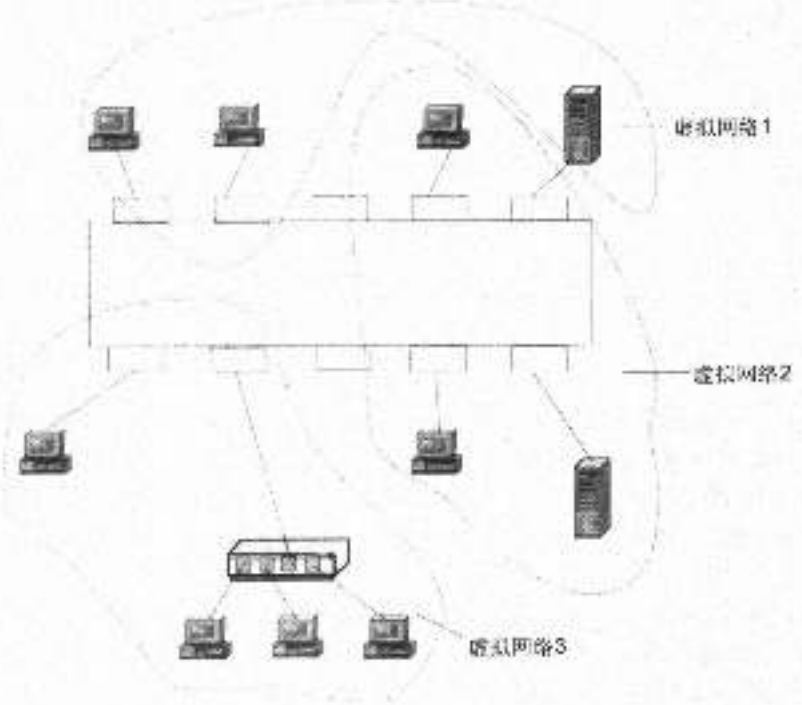


图16-9　基于一个交换机建立的VLAN

说明　如果一个基于中继器的网段连接到交换机的一个端口，则将这个网段的节点分给不同的VLAN是没有任何意义的，任何情况下，这些节点的流量仍然是公共的。

基于端口聚合而建立的虚拟网络并不需要管理员大量的手工操作——只需把每个端口赋给一个事先命名了的VLAN就足够了。通常，这个操作是由交换机提供的特殊程序执行的。

建立VLAN的第二种方法是基于台址MAC地址。每一个由交换机学习到的MAC地址都被赋给一个VLAN。当网络由许多节点组成时，这种方法就要求管理员进行大量的手工操作。然而，当在多个交换机的基础上生成VLAN时，这种方法要比端口聚合方法更灵活。

16.4.3　构建基于多个交换机的VLAN

图16-10给我们展示了当在多个支持端口聚合（port grouping）技术的交换机的基础上构建VLAN时会产生的问题。

如果某些VLAN的节点可以横越多个交换机，那么必须为交换机指定一对特殊的端口用于连接每个VLAN的交换机。否则，当交换机仅仅由一对端口相连时，在交换机与交换机之间传送帧的时候，某个从属于哪个VLAN的信息就会丢失。因此，为了获得连接，具有端口聚合的交换机要求端口的数目与支持的VLAN的数目相一致。但是，当使用这种方法时，端口和电缆会很浪费。不仅如此，当使用路由器连接VLAN时，要为每一个VLAN分配一个独立的网线和独立的路由器端口，这

也需要可观的管理费用。



图16-10　在具有端口聚合的多个交换机的基础上建立VLAN

在每个交换机上，把MAC地址聚合成一个VLAN，消除了使用多端口连接交换机的必要，因为在这种情况下，MAC地址表示一个VLAN标记。然而，这种方法需要大量的手工操作来将MAC地址映射到每个网络交换机上的VLAN上。

这里讨论的两种方法都是仅基于在交换机地址表上增加的一些辅助信息。它们缺少直接在地址表中插入某个帧属于的哪一个VLAN的帧信息的可能性。当帧在交换机之间传送的期间里，剩下的方法使用帧已有的或者辅助的域来存储帧的VLAN信息。此时，就没有必要在每个网络交换机上保存MAC地址与VLAN的映射关系。

只有当帧在交换机与交换机之间传送时，才会使用一个包含特定VLAN标记的附加域。当把帧传送给端节点时，通常会删除这个附加域。这样，这种“交换机—交换机”交互协议就被修改了，而端节点的软硬件仍保持不变。在IEEE 802.1Q标准被采用前，有很多这种类型的专用协议。然而，所有的这些协议都有一个共同的缺点——当建立VLAN时，不同生产厂家的设备互不兼容。

为了存储虚拟网络的网络号，IEEE 802.1Q标准（也被称做标签协议，因为它给头部增加标签。）提供了一个附加的4字节的头部（图16-11）。头部的前两个字节组成了**标签协议标识符**（Tag Protocol Identifier，TPID），常常是一个16进制数0x8100。这样网络设备就可以识别这个帧是带有标签的以太网帧。紧接着的两个字节叫做**标签控制信息**（Tag Control Information，TCI）。802.1Q协议与802.1p协议共享这几个字节，这些我们将在16.5节中详细讨论。在这个域中，有12比特用于存储VLAN号（VLAN ID域），3比特用于存储在802.1p标准中定义的帧的优先级（用户优先级域），1比特，被称为**规则格式标识符**（Canonical Format Identify，CFI），用来区别以太网帧和令牌环帧。对于以太网帧而言这一位结果为0。12比特的VLAN ID域最多允许建立4 096个虚拟网络。因为，为了添加802.1Q/p头部，以太网帧的数据域减少了2字节，所以它的最大大小也减少了。例如，对于非以太网的帧来说，这个大小为42~1 496字节，而标准值是46~1 500字节。

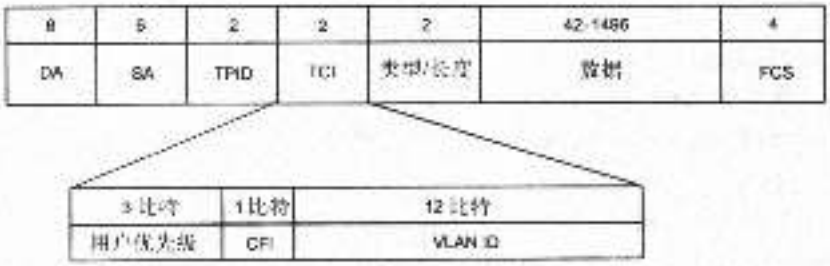


图16-11　带有标签的以太网帧的结构

802.1Q标准的采用，使得设备制造商可以克服专用VLAN实现的差异，从而实现了建立VLAN时的兼容性。交换机的生产商和网络适配器的生产商都支持VLAN技术。在后一种情况下，适配器

可以产生并接收包含VLAN TAG域的标签以太网帧。如果网络适配器产生标签帧，它就将这些帧与一个特定的VLAN相联系。因此，交换机就必须适当地处理这些帧（例如，决定是否需要将这个帧传送给某一个特定的输出端口），这取决于与特定VLAN映射的那个端口。网络适配器驱动程序从网络管理员手工输入的配置数据那获得它的VLAN或VLAN的编号；另一方面，它也可以从运行在某个特殊节点上的某个应用程序那里收到这些信息，这个应用程序也可以在网络中心服务器上运行，并控制整个网络的结构。

网络中VLAN的存在影响了活动生成树拓扑的选择。考虑图16-12所示的这个例子。

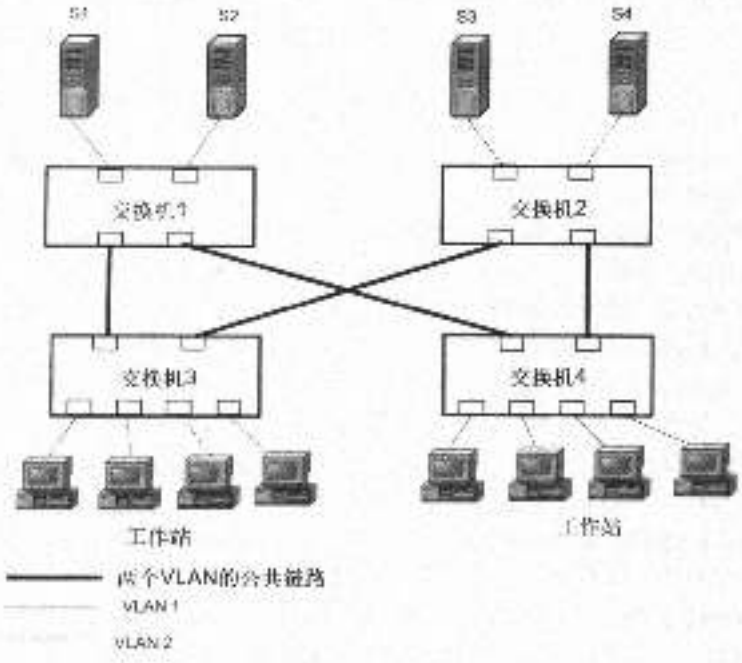


图16-12　具有VLAN和冗余链路的网络

在这个网络中有两个VLAN：VLAN 1和VLAN 2。这些VLAN建立在端口聚合技术的基础上。在这个图中，我们用实线来表示VLAN 1端口间的连接，用虚线来表示VLAN 2端口间的连接。

当建立活动STA拓扑的时候，我们没有考虑网络中VLAN的存在，结果就是图16-13中所示的生成树（交换机1被选为根交换机）。这种拓扑对于VLAN 2而言效率不高。例如，从计算机W1到服务器S3需经过4个转发交换机。相对的，从VLAN1的计算机到服务器S1只需要通过2个转发交换机，但是如果交换机2被选为转发交换机，

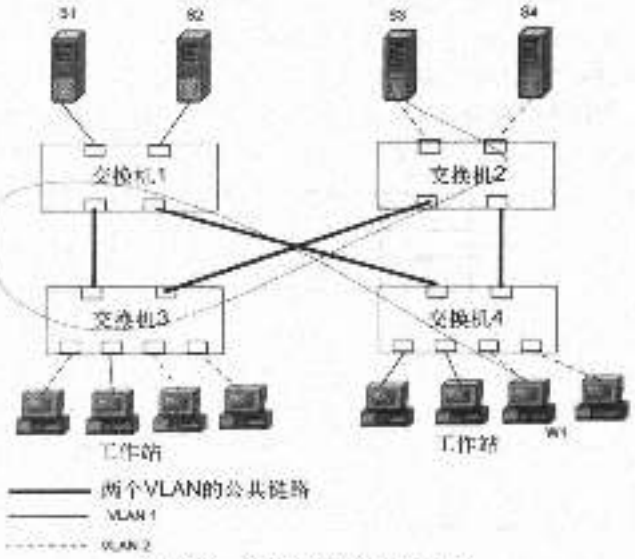


图16-13　不考虑VLAN的生成树

那么VLAN1的效率就不高了。

另一种解决方法是分别为每一个VLAN建立活动的拓扑。在这个例子中，这种方法将会产生两棵树，为VLAN 1选择交换机1作为根交换机，为VLAN 2选择交换机2作为根交换机（图16-14）。

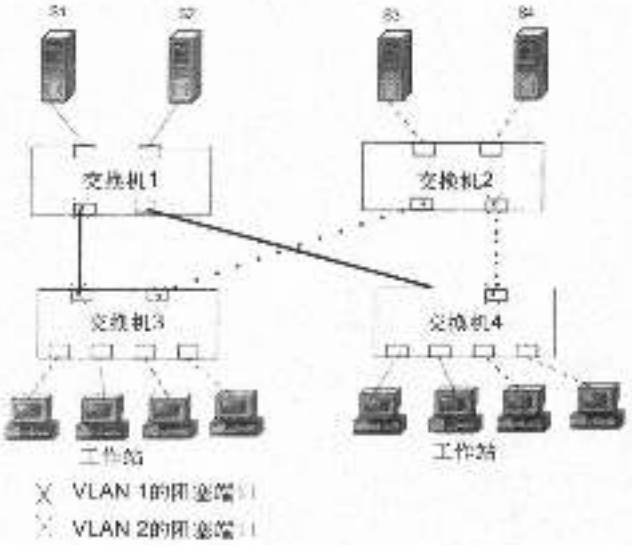


图16-14　考虑VLAN的生成树

16.5　LAN中的服务质量

LAN交换机几乎支持第7章中描述的所有的QoS机制。当LAN 交换机作为一种通信设备时，这个论断总是正确的。另一方面，每个交换机模型都可能仅支持一小类特定的QoS机制，或者不支持这些机制。通常，工作组交换机不支持QoS，但是对于主干交换机，这个支持是必须的。

流量分类。LAN交换机是第二层设备，它只分析数据链路层的头信息。因此，交换机通常使用源和目的MAC地址以及收到达的端口号来对流量进行分类。它也可以用来对数据域中的任意子域进行分类，这些子域由字节偏移量来指定。这些方法对于管理员来说并不方便，例如，他们需要把语音流量从文件传输流量中分离出来，因此，一些交换模块仅基于包含在的高层协议的头部中的特征值来分类，而没有完全地支持它们（例如，没有使用IP进行分组转发）。例如，这种分类可以基于包含在分组头部中的IP地址和应用特征值来执行。

流量标签。流量标签仅仅在网络边缘进行分类，然后将分类的结果用于网络中的所有转发设备。以太网802.3帧并不包含任何可以存储流量分类结果的域。然而，这种不足在802.1p标准中得到了修改，它使用3比特，也就是前面提到的802.1Q/p附加头部中用来存储帧优先级的那3个比特。

这3比特用于存储8种可能的流量分类中的一种，包含了802.1p规范的802.1D-1998标准就是这样来解释这个域的。附录H中的802.1D-1998标准提供了将所有的LAN流量分成8类的建议，表16-1中列出了这8个分类。

表16-1　LAN流量分类

用户优先级	缩写	流量类型
1	BK	背景
2		无
0（默认）	BE	最大努力

（续）

用户优先级	缩写	流量类型
3	EE	极好努力
4	CL	受控制的负载
5	VI	“视频”，<100ms延迟和抖动
6	VO	“音频”，<100ms延迟和抖动
7	NC	网络控制

背景（BK）是对延迟最不敏感的流量，例如备用流量。然而，这个流量的源可能会传送可观的数据量。因此，将其分为一个单独的类别是有意义的。这保证了这种流量不会减缓其他类型流量的处理。

最大努力（BE）、极好努力（EE）以及受控制的负载（CL）这几类并不是实时的类型，这表示它们不会对延迟极限强加苛刻的要求。然而，对于这些类，它需要保证某个最小级别的带宽。对于这些类使用加权队列机制是很适宜的。

视频（VI）、音频（VO）和网络控制（NC）这几类是延迟敏感的。表16-1提供了延迟阈值的建议值。对于这几类使用优先级队列机制是比较好的方法。网络控制类优先级最高，这是由于所有的网络特征依赖于实时的决策和传输到网络设备的信息。

队列管理。支持QoS的交换机使用多个队列来区分不同类型流量的处理。这些队列可以是优先级队列，可以是加权队列，也可以兼而有之。

通常，交换机支持的队列数量是有一个最大值的，这个值可能小于需要的流量类型值。在这种情况下，一些类可能要共用同一个队列，也就是说它们必须合并成一个类。802.1D-1998标准给出了下列建议，这个建议给出了在队列数目一定的条件下网络必须实现的流量类型（表16-2）。

表16-2　流量分类和队列数

队列数目	定义流量类型							
1	BE							
2	BE				VO			
3	BE				CL		VO	
4	BK		BE		CL		VO	
5	BK		BE		CL	VI	VO	
6	BK		BE	EE	CL	VI	VO	
7	BK		BE	EE	CL	VI	VO	NC
8	BK	—	BE	EE	CL	VI	VO	NC

当网络中只有一个队列的时候，网络中唯一可以存在的流量类型是BE。它的QoS不能通过队列管理来提升，虽然如反馈和带宽预留这样的能力仍然是可用的。

如果网络中有两个队列，流量就可以被分成两类，BE和VO。在这种情况下，所有延迟敏感流量都应被归类为VO，这些流量中不仅包括音频流量，还包括视频流量和网络管理流量。

更多的队列数允许更多不同的流量服务，直到类的数目增加到所推荐的8个。

上面提到的方法只是一个建议，网络管理员可以按他们的需要进行分类。

此外，服务单独的流量流也是可能的。然而，在这种情况下，每个交换机都必须独立的将聚合流量分离成独立的流，因为以太网帧没有专门的域用来存储流标记。

VLAN的编号也可以用作流量类型的特征值。这个特征值可以与帧优先值域相结合，这样管理

员就可以产生很多不同的流量类型。

预留和监管。LAN交换机支持为流量类型或者个别的流保留接口带宽。通常，交换机允许管理员给这个类型的流量或者流分配某个最小信息速率，这是用于拥塞时期的保障速率；管理员还可以分配最大信息速率，这由监管机制来控制。

对于LAN交换机而言，没有标准的资源预留协议。因此，为了执行这些预留，网络管理员必须单独地配置每一个网络交换机。

16.6 网桥和交换机的局限性

交换机的使用使得网络管理员可以克服一些限制，典型的有基于共享介质的网络的限制。交换机网络可以横跨相当大的地域，平稳地变成MAN。它们也可以包含具有不同带宽的网段，这样就组成了高性能的网络。最后，它们可以使用可替换的路由来提升网络性能和可靠性。

然而，仅仅基于中继器、网桥和交换机（例如，没有使用网络层设备）来建立复杂的网络具有显著的局限性和缺点：

- 第一，交换LAN拓扑仍然具有相当大的局限性。STA技术和链路聚合技术的使用，消除了规定网络必须部分的无环路这个要求所强加的限制。STA不允许可替换路由用于传送用户数据，而链路聚合技术也仅允许在两邻居交换机之间的网段中这么使用可替换路由。这些限制不允许许多高效的拓扑，而这些拓扑本可以用于流量传输的。
- 第二，位于交换机间的逻辑网段之间是弱分离的，这表示它们缺少抵御广播风暴的能力。虽然大多数的交换机实现的是VLAN机制，并且这个机制允许灵活地创建不共享流量的独立的工作站组，但是这种解决方案并不是没有缺点的。VLAN机制隔离了VLAN个体，从而使得一个VLAN中的节点不能与另一个VLAN中的节点交互。
- 第三，在基于网桥和交换机的网络中，没有简单的方法可以解决基于分组中的数据值的流量过滤问题。在这种网络中，这项任务只有使用用户自定义的过滤器来完成。而为了建立用户自定义的过滤器，网络管理员必须处理二进制形式的分组内容。
- 第四，仅使用交换机支持的物理层和数据链路层的工具的运输子系统的实现导致了一个不够灵活的平面寻址系统：MAC地址被硬编码到网络适配器中作为目的地址。
- 最后，在建立异构网络时，交换机只具有有限的协议转换能力。例如，由于这些网络中寻址系统的不同、以及数据域的最大值的不同，交换机不能将WAN协议转换成LAN协议。

上面提到的数据链路层协议的这些严重的局限性使得仅仅基于数据链路层工具建立大型的异构网络是问题多多。解决这些问题的最自然的方法就是使用更高一层的——网络层的工具。

16.7 案例学习

在第12章中，我们考虑了建立在10Mb/s以太网中继器基础上的LAN结构，这个结构用于满足Transmash工程工厂的内部需要。这个例子对应的情形在20世纪90年代早期很具有代表性，那时一个10Mb/s共享介质就可以完全满足整个企业的需要，它可以允许企业各部门的计算机间的流量交换，当然，数目很少。这里，我们考虑这个工程工厂的LAN的一个升级版本，它在90年代后期成为大多数大型企业网络的典型。

这种网络的一个主要特点就是整个LAN都是建立在交换机的基础上的（图16-15）。移植到交换网络主要是由于对于LAN的性能和可靠性的进一步的要求。在那之前，计算机数据处理对生产过程至关重要。因此，计算机的数量也急剧上升，还引进了新程序，这样就可以交换大量的多媒体信息。

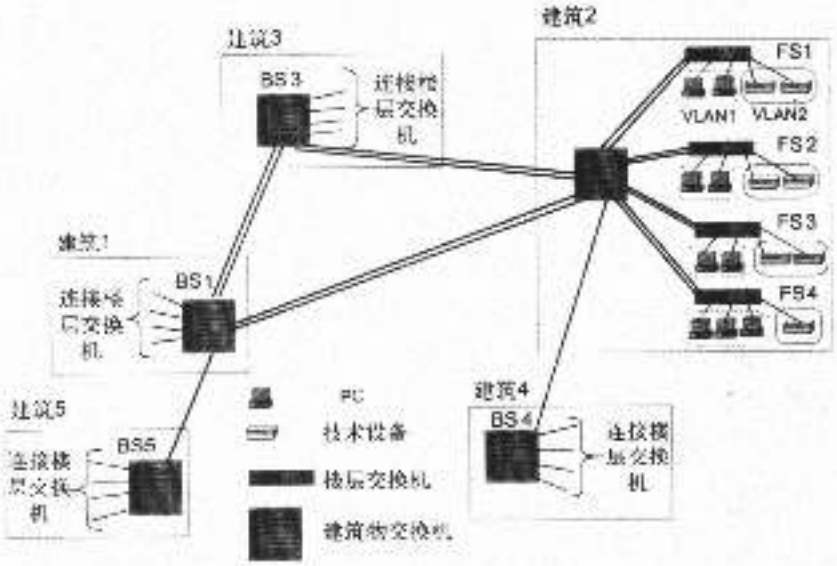


图16-15　Transmash工程1.1的交换网络

这几个建筑中的任意一个的LAN的基础都是由一个强大的中心交换机组成的，这个交换机有一个机箱，装备有快速以太网端口和千兆以太网端口（交换机BS1-BS5）。建筑物交换机连接楼层交换机，这些楼层交换机使用由两到三个快速以太网端口组合成的干线连接到建筑交换机。楼层交换机用于连接两种端用户设备：PC和多种技术设备（这些连接的细节可以参见例子中建筑2的LAN）。PC用户使用形成自动化企业资源计划系统（ERP）的应用程序，技术设备使用形成自动化计算机辅助制造（CAM）系统的应用程序。

建筑物2、建筑物3和建筑物4的中心交换机组成了企业LAN的主干。它们由2个端口的千兆以太网干线连接，保证了可观的性能储备。建筑物4和建筑物5使用普通的千兆以太网（没有使用干线）连接到主干上。为了连接这些建筑物的交换机，使用了多模光纤。这种光纤是为了基于中继器的LAN而安装的，因为它的质量可以用来保证1 000Base-SX端口操作的稳定性。

Transmash网络传送两种流量：分别是来自ERP应用和CAM应用的流量。这两种流量的区别在于它们的QoS要求不同。例如，CAM流量是实时流量，而ERP流量不是。因此，Transmash网络中组织了两个VLAN——VLAN 1用于ERP流量，VLAN 2用于CAM流量。这使得可以对每种流量进行可靠的分离，并附加地简化了交换机所支持的QoS，因为VLAN号（在这个例子中是2）表明了流量必须在优先级队列中进行处理。

由于VLAN的主干具有冗余链路，所以交换机为每个VLAN单独使用一个STA。对VLAN 1而言，BS4和BS3交换机之间的链路是预留链路，VLAN 2的预留链路是BS3与BS4交换机之间的那条。ERP和CAM之间可以进行数据交换，因为这两个VLAN有多个服务器成员。

小结

- 为了在复杂网络中自动支持预留链路，交换机使用了生成树算法（STA）。对这个算法的描述请见IEEE 802.1D标准。STA基于周期性地在交换机之间交换特殊的帧。通过这些帧，交换机找到并且阻塞网络中存在的闭合回路。
- STA协议通过3个步骤发现生成树结构。第一步，确定根交换机；第二步，确定根端口；第三步，选择网段的指定端口。

- STA 802.1D协议的主要不足是需要相对较长的时间来建立一个新的活动结构——大约50秒。更新的802.1w标准改正了这一不足。
- 把多个物理链路聚合成一个逻辑信道是在基于交换机的LAN中使用一些可替换路由的一种形式。
- 链路聚合同时提升了网络可靠性和性能。
- 不仅可以在两个相邻的交换机之间建立聚合信道，而且还可以在多个交换机的端口之间也可以建立聚合信道。为了自动通知某个物理端口属于哪个聚合端口，开发了一个特殊协议，这个协议被称为链路聚合协议。
- 虚拟LAN（VLAN）技术允许在基于交换机的网络中建立端节点的独立分组。这些独立的分组之间没有包括广播流量在内的一切流量。
- VLAN结构常常由端口聚合生成。为了建立一个基于多个交换机的VLAN，需要用一个特定的标签标记这些传送的帧，这个标签标识了这个帧的发送端所属的网络号。
- 在802.1Q规范中定义了VLAN标签的标准格式。
- LAN交换机支持所有类型的QoS机制：流量分类和监管、优先级队列和加权队列、以及带宽预留。

复习题

1. STA的目标是什么？
2. 生成树的定义是什么？
3. 交换机的哪一个端口叫做根端口？
4. 以下定义中的哪个描述的是指定端口：
   A. 网络管理员根据自己的判断指定的根端口
   B. 在特定网段中，与根交换机距离最短的端口
   C. 转换到阻塞状态的端口
5. STA中，交换机之间的距离是如何度量的？
6. 列出建立动态生成树的3个过程。
7. 如果与根交换机的距离都相等，如何从几个候选端口中选择根端口？
8. 网络管理员可以影响对根交换机的选择吗？
9. 交换机如何确定选择动态拓扑的过程完成了？
10. 什么触发交换机去选择一个新的拓扑？
11. STA的主要缺点是什么？
12. 链路聚合：
    A. 提升了网络性能
    B. 提高了网络的可靠性
    C. 两者皆有
13. 在什么情况下使用链路聚合比移植到更快版本的以太网技术更有效？
14. STA和链路聚合如何交互？
15. 链路聚合技术的局限性有哪些？
16. 单向干线和双向干线的区别是什么？
17. 当选择用于帧传输的干线端口时，需要考虑些什么？
18. 当使用聚合链路时，为什么需要考虑属于相同会话的帧？
19. 为什么VLAN可以称为广播域？

20. 如何使连接多个VLAN成为可能?
21. 列举建立VLAN的主要方法。
22. 为什么在基于多个交换机的网络中端口聚合的效率不高?
23. 在IEEE 802.1Q标准中，选择的是哪个方法来解决在多个交换机的基础上建立VLAN的问题?
24. 可以同时使用端口聚合和IEEE 802.1标准吗?
25. STA是否需要考虑网络中是否存在VLAN吗?
26. LAN交换机支持QoS的哪一种机制?
27. IEEE 802.1D-1998推荐的流量类型有多少?
28. 如果网络交换机支持的队列数小于流量的类型，你将怎么办?
29. 列出基于交换机建立的网络的局限性。

第四部分　TCP/IP网际互联

在学习了本书中的大部分内容之后，请回顾一下前面三个部分所学到的知识，并思考接下来两个部分我们将学些什么。在第一部分中，仅从概念层次描述了本书中的问题，这可能是书中最复杂和重要的部分。事实，学识水准和专业技能水平都在很大程度上依赖于它们所搭建的框架。我们已经多次引用第一部分中的材料，接下来还会继续这么做。

第二部分和第三部分致力于在物理层和数据链路层数据传输的特定技术，几乎没有涉及到以图表或计算机"飘浮"在其中的"云朵"形式的抽象网络模型。相反，却介绍了一些特定协议、帧格式以及网络设备。

本书的第四部分会介绍些什么主题呢？根据OSI模型逻辑所提示的，考虑物理层和数据链路层技术之后必须介绍网络层工具，这些工具可以将多个网络内联成一个大型网络。因为IP在所有网络层协议中起着毋庸置疑的领导作用，我们将以此协议为示例来考虑内联网的所有特性。然而，由于IP和TCP/IP栈其他协议之间的紧密联系，我们会尝试着提供一种宽泛的模式来解释它们之间的交互。

注意到前面几章，我们曾经提及甚至解释过直接与TCP/IP内联网络相关的不同方面。在第2章中，我们考虑了路由的基本概念和原则。第4章与OSI模型的网络层信息一起介绍了内联网的概念。根据那里提供的定义，通常说的网络是若干个网络的组合，称为内联网（internetwork）或因特网（internet）。组成内联网的网络称作子网、组成网或者只是网络。子网通过路由器相连。因特网的成员可以是LAN和WAN。每个组成网络中的所有节点之间利用相同的技术进行通信，譬如以太网、令牌环、FDDI、蜂中继或X.25。然而，要想创建两个任意选择、属于不同网络的节点之间的连接，这些技术中没有一个能够做到。这项任务，也就是中任意两个节点之间的交互，可以由TCP/IP栈协议来完成。在第5章中，我们提供了对因特网结构的描述，这是基于TCP/IP技术所建造的最大型的网络。我们强烈推荐读者复习这部分内容。

在本书的最后部分，与各种WAN技术相关的部分，我们又回到了TCP/IP。我们将考虑ATM/FR上构建IP的特殊特性，以及与IP紧密相关的多协议标签交换技术。同时还包括简单网络管理协议和IP的安全版本-IPSec。

第17章 TCP/IP网络中的寻址

17.1 引言

TCP/IP技术旨在解决以下寻址问题：

- 协调使用不同类型地址（*Coordinating the use of different types of address*）。这包括不同地址间的映射，如将网络IP地址转换为本地地址或者将域名映射到特定IP地址。
- 确保地址的唯一性（*Ensuring address uniqueness*）。根据地址类型，需要在特定的计算机、子网、互联网或企业网或因特网中确保地址的唯一性。
- 配置网络接口和网络应用程序（*Configuring network interfaces and network application*）。

对于十个节点的小型网络来说，以上每个问题都有非常简单的解决方案。比如，要想把一个符号域名映射到特定的IP地址，每个主机上都保存域名到IP地址的映射表就足够了。在小型网络里手动指定特定地址到所有接口的映射也不是很难。然而，大型网络中，这些任务变得非常复杂从而需要本质上完全不同的解决方案。

于是，网络的可延成为标志这些问题解决方案的关键字。

TCP/IP提供的IP地址指派、映射和配置步骤在不同规模的网络中表现同样出色。本章中，除了IP寻址以外，还包括最流行的可延拓的工具来确保对TCP/IP网络寻址的支持：无类别域间路由（CIDR）、域名系统（DNS）以及动态主机配置协议（DHCP）。

17.2 TCP/IP栈的地址类型

为了区分不同的网络接口，TCP/IP网络使用下面三种地址类型：

- 本地（硬件）地址
- 网络（IP）地址
- 符号地址（域名）

17.2.1 本地地址

大多数LAN技术，如以太网、FDDI以及令牌环使用**MAC地址（MAC addresses）**定义接口，也有一些其他技术（比如，X.25、ATM和帧中继）使用不同的寻址机制，这些机制在相关技术组建的网络中同样使用唯一地址。这种自治的网络使用排它的寻址系统来实现内部目标——以保证节点的连接性。然而，一旦某个网络被连接到其他网络，这些地址的功能就被延伸了，它们就成为更高层网间技术的必备元素——这就是TCP/IP技术。TCP/IP中这些地址所扮演的角色并不依赖其组成网络中的特定技术。因此，这些地址都有着共同的名字：**本地（硬件）地址（local / hardware addresses）**。

> **注意** 这里使用的定义——本地和硬件——可以做另一种解释。“本地”在TCP/IP语境中意味着该地址在组成网络而非在整个互联网中有效，这里有必要解释下术语：本地技术（建立组成网络的技术）和本地地址（本地技术用来在组成网络中为节点寻址）。回忆一下组成（本地）网络可以同时建立在WAN技术（X.25、帧中继等）和LAN技术（以太网、FDDI等）上。术语“本地”也被用于LAN——局域网中。然而，在这里它有着不同的含义，代表着将网络局限在小范围内的技术特征。

在网际间环境中解释"硬件"也有一定的难度。在这种情况下，它强调TCP/IP设计开发者将组成网络解释成一个辅助的硬件工具，其唯一目的是利用组成网络向最近的路由器发送IP分组。底层的网络技术可能很复杂，但是既然TCP/IP技术并不牵涉到这些细节，所以这也无关紧要。

举个例子，考虑某TCP/IP互联网中包含一个采用IPX/SPX网络的情况，后者可能被拆成若干子网，与IP网络相似的是，它也用硬件地址和IPX网络地址来表示节点。然而，TCP/IP网络忽视IPX/SPX网络的多层结构，认为它是普通的组成网络。这一点与以太网处理方式相同。因此，TCP/IP技术将IPX/SPX网络中的网络地址看做本地地址。同样，如果组成网络建立在X.25技术的基础上，那么X.25地址也会被IP技术看做是本地地址。

17.2.2 IP网络地址

为了完成网络互联的任务，TCP/IP技术需要拥有自己的全局寻址系统，并且这个寻址系统不依赖于组成网络中的节点寻址方法。该寻址系统必须提供统一的方法来唯一地标识互联网上的每个接口。

有一种构成网络地址的自然方式就是唯一标识所有的组成网络，并为每个网络中的节点标号。这样，**网络地址**（network address）就由一对号码组成：一个**网络号**（network number）和一个**节点号**（node number）。

节点号可能这样组成：该节点的本地地址（IPX/SPX技术采用的方法）或一些与本地技术无关但能唯一标识子网中该节点的数字串。

第一种方法中，网络地址依赖于本地技术。这限制了它的适用领域。比如，IPX/SPX网络地址被设计成使用在某种特定的互联网中，该互联网还接只使用MAC地址或者类似格式的地址的网络。第二种方法则适用得多，这正是TCP/IP栈的特征。

在TCP/IP技术中，网络地址又被称为IP地址（IP address）。

注意 考虑IP网络环境。路由器，根据其定义，处在多个网络中。这样，每个接口都应该有自己的IP地址。每个节点同样可以加入到多个IP网络中。这种情况下，计算机必须为这些网络连接准备足够多的IP地址。因此，IP地址标识网络连接而不是特定的计算机或路由器。

当分组通过互联网传送给接收端时，头部必须含有目的节点的IP地址。每个路由器通过目的网络号找到下一个路由器的地址。在将分组传给下一个网络之前，基于在下一个路由器IP地址，路由器必须判定它的本地地址。IP使用地址解析协议（ARP）来实现这个目标（见图17-1）

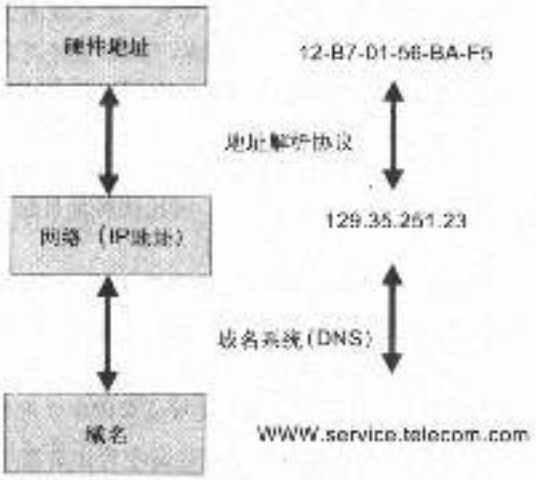


图17-1 地址解析协议

17.2.3 域名

TCP/IP网络的硬件和软件根据IP地址识别计算机。比如，FTP://192.45.66.17命令建立与该FTP服务器的连接，而HTTP://203.23.106.33命令打开该网络服务器上的主页。然而，多数用户倾向于使用计算机的符号名称，因此TCP/IP网络必须为主机实施符号名称并且提供将其映射到IP地址的机制。

互联网中网络接口的符号标识符依据层次原则建立。IP网中的完全合格符号或域名的各个组成部分用小数点分开，以下列顺序排列：个人主机名称、主机组的名称（如组织名称）、更大的组名称（域），依此类推至最高层的域，比如根据地理位置定义的机构：us-美国，ru-俄罗斯，uk-英国。域名应该如下所示：server2.janet.ac.uk。

域名和主机的IP地址之间不存在功能性依赖关系，因此符号名称映射到IP地址的唯一方法就是使用表格。TCP/IP网络使用特殊的分布服务，域名系统（DNS）由网络管理员建立一张映射表实现上述映射。因此，域名又被称为**DNS名称**（DNS names）。

总之，网络接口可以同时拥有一个或多个本地地址、一个或多个网络地址以及一个或多个域名。

17.3　IP地址格式

IP分组的头部有两个域用来存储发送端和接收端的IP地址，每个域固定长度为4个字节（32比特）。IP地址由两个逻辑组成部分构成——网络号和该网络中的主机号。

最流行的IP地址的表达方式是4个数字的形式，用十进制表示法表示每个字节的值每个数字之间用小数点分开。典型的点式表达的IP地址如下所示：128.10.2.30

同样地址的二进制表示如下：10000000 00001010 00000010 00011110

用十六进制表达：80 0A 02 1D

注意，该地址形式中并没有网络号和主机号的明显界限，然而当分组在网络上传输时，经常需要将地址分成两个部分。比如，根据规则，路由是基于网络号执行的，每个收到分组的路由器必须搜索头部的特定区域来找到目的主机地址然后将网络号放在该地址内。路由器怎样知道32比特的IP地址中哪些是网络号、哪些是主机号呢？

下面来看几种解决办法。

- 最简单的解决方法是使用**固定边界**（fixed boundary）。这种情况下，整个32位被提前拆分成两部分。这两部分不等长，但必须固定长。一部分总是包含网络号，另一部分用来存储主机号。这个方法非常简单，然而它好用么？不是非常好。因为分配用来存储主机号的域长始终相同，所有的网络都拥有同样的最大节点数。比如，假设只分配第一个字节存储网络号。这种情况下，整个网络空间被拆成相对少量（2^8）的大型子网（每个包含2^{24}个主机）。若把分界线往右挪，将会有更多的相同大小的子网。显然，这样严格的解决方案无法满足不同组织结构的差异性需求。因此这种地址结构虽然在RFC760中被定义为TCP/IP改革的第一步，但没有得到广泛的应用。
- 第二种方案（RFC950和RFC1518）使用掩码来解决。对于建立网络号和主机号之间界线，掩码是最灵活的方法。用这种方法，整个地址空间可以被表示成不同大小的一系列子网。

掩码代表着与IP地址同时使用的一串数字，二进制掩码包含的一串数位1代表着网络号的位置。掩码中1字串和0字串的边界就对应着IP地址中的网络号与主机号的分界点。

- 最后，第三种方法，也是迄今为止最常用的，利用RFC791中定义的**IP地址分类**（IP address classes）。这个方法综合了前面描述的两个方法：尽管子网大小不是任意变化的，如使用掩码时所示，但它们也不是一成不变的，如设定固定边界时所示。本方法定义了地址的五个分类，其中三个用于网络寻址，剩下两个被保留以备特殊用途。

17.3.1　IP地址的分类

地址的前几个位值用来做IP地址分类的准则，表17-1列出了不同地址分类的IP地址结构。

表17-1 IP地址的分类

类别	第一字节	最小网络号码	最大网络号码	节点数量
A	0	1.0.0.0（0-不被使用）	126.0.0.0（127-保留）	2^{24}（3字节）
B	10	128.0.0.0	191.255.0.0	2^{16}（2字节）
C	110	192.0.0.0	223.255.255.0	2^{8}（1字节）
D	1110	224.0.0.0	239.255.255.255	多播地址
E	11110	240.0.0.0	247.255.255.255	保留

- **A类地址**（Class A）中的最高有效位的值为0。其中1个字节分配给网络号，另外3字节分配给子网中的主机号。首字节在1（00000001）到126（01111110）范围内的IP地址都属于A类地址，网络号0（00000000）不使用，网络号127（01111111）保留以备特殊用途，这个本章后面会详细讨论。A类网络并不是无穷尽的，然而，其中主机的数目可以多达2^{24}（也就是16 777 216）个。
- **B类地址**（Class B）前两个最高有效位的值为10，其中，2个字节分配给网络号，2个字节分配给主机号。前两个字节的范围在128.0（10000000 00000000）到191.255（10111111 11111111）内的IP地址都属于B类地址。很自然，B类网络比A类网络数目更多，但是规模要小得多。B类网络中的最大主机数为2^{16}（也就是65 536）个。
- **C类地址**（Class C）前三个最高有效位的值为110，其中3个字节分配给网络号，只有1个字节分配给主机号。前三个字节的范围在192.0.0（11000000 00000000 00000000）到223.255.255（11011111 11111111 11111111）内的IP地址都属于C类地址。C类网络使用最为广泛，其主机数目被局限为2^{8}（256）个。
- 如果IP地址以1110序列开头，那么它属于**D类地址**（Class D）并且属于特殊的组地址。与识别特殊网络接口的A类、B类、C类相比（比如，**单播**（unicast）地址），组地址又被称为**多播地址**（multicast address）。组地址识别的网络接口组通常属于不同的网络，每个组的接口除了自身的IP地址外，还被指派一个组地址。传送分组时，当D类地址被指定为目的地址，那么该分组将被传送到组中的所有主机。
- 如果IP地址以11110序列开头，那么它属于**E类地址**（Class E）。这种地址被保留以供日后使用。

说明 为了从IP地址中获得网络号和主机号，需要将地址分为两部分然后用0补充每一部分成为4个字节。比如，假设有下列B类地址：129.64.134.5，前两个字节识别网络号，后面两个指定主机号。因此，网络号为129.64.0.0，主机号为0.0.134.5。

17.3.2 特殊的IP地址

TCP/IP分配IP地址时有一个限制，就是网络号和主机号不能由全1或全0组成。因此，表17-1中各类网络的最大主机数减少2个，如C类网络中8位用以存储主机号，这样总共有256个号可以使用：从0到255。但事实上，C类网络中最大主机号不会超过254，因为0和255不能用做主机号。出于同样的考虑，端节点的IP地址不能有类似于98.255.255.255这样的IP地址，因为在这个A类网络中，主机号（0.255.255.255）只包含二进制1值。

因此，IP头部中的一些IP地址用以下方式解释：

- 如果IP地址由全0组成，称为**未定义的地址**（undefined address）并指向产生该分组的主机地址。特殊情况下，这种类型的地址被放在IP分组头部的源地址域中。
- 如果网络号由全0组成，默认条件下目的主机与发送分组的主机属于同一子网。这种地址只

可以用作源地址。

- 如果IP地址为全1，具有这样的目的地址的分组必须被传送到与发送分组同一个子网中的所有主机。这种传送方式被称为**有限广播**（limited broadcast），“有限”意味着分组永远不会离开该子网。
- 如果目的地址的主机号为全1，这样的分组被送往同一子网中的所有主机。比如，目的地址为192.190.21.255的分组会被送往192.190.21.0子网中**所有**（all）主机。这种传送方式又叫做**广播**（broadcast）。

注意　IP的多播概念与数据链路层的LAN协议中的不同，后者数据毫无例外地被送往所有主机。在互联网中，IP有限广播和IP广播都有传播限制，它们或者被源主机所属的网络边界所限制，或者局限于目的地址所在的网络中。因此，利用路由器的网络分组局限在某个子网范围内来进行广播，只是因为没有办法对互联网中的所有主机同时进行寻址。

IP地址中第一个字节被设成127时有着特殊的意义。该地址是计算机和路由协议栈的内部地址，用来测试程序或协调安装在同一台机器上应用程序的客户端组件与服务器端组件之间的合作，而且两个组件之间存在着网络通信。然而，客户端与服务器端共存在于同一台机器上，应该使用怎样的IP地址呢？可以用主机的网络接口地址。这样的话，网络中将不可避免地出现冗余分组；因此这个方案不能解决问题。使用127.0.0.0内部地址更加经济有效。或许你还记得，以127开头的IP地址不允许用作网络接口地址。当一个程序往类似127.x.x.x的地址发送数据时，该数据并不会被送到网络上。取而代之的是，这样的分组被返回到上一层网络协议，此类分组的路由是一个循环，因此该地址被称为**回环**（loopback）。

属于D类的**多播**（multicast）地址旨在提供一个经济有效的方法将语音和视频程序传送给因特网或企业网上广大的接收端。如果IP分组的目的地址域存放的是多播地址，它必须被传送到用地址域中指定的数组成组的多个节点。同一主机可以属于多个组，特定多播组的成员也没必要必须属于同一个子网。总的说来，它们可以分布在任意间距的不同子网上，多播地址无需区分网络号和主机号，路由器会用特殊的方式来处理。

多播地址的主要目的是“一对多”地发布信息。当前多播地址只用在小型的、试验性质的网络范围内，好比大洋中的一个个孤立的小岛。因特网是否能变成广播和电视的有力竞争者，还要依赖于多播机制是否能得到广泛应用。

17.3.3　在IP地址中使用掩码

将IP地址与掩码结合使用就可以舍弃地址分类的概念，使得寻址系统更加灵活。

回想一下，对于标准的网络分类，子网掩码可以如下取值：

A类	11111111.00000000.00000000.00000000	(255.0.0.0)
B类	11111111.11111111.00000000.00000000	(255.255.0.0)
C类	11111111.11111111.11111111.00000000	(255.255.255.0)

这种解决方案的主要思想是利用子网掩码，其决定网络号边界的二进制码的位数不一定是8的倍数（也就是说，地址被分成一个个字节）。比如，如果你用子网掩码255.255.255.0映射185.23.44.206，那么网络号是185.23.44.0，而不是如分类系统定义的185.23.0.0。

考虑另外个例子：假设255.255.128.0为IP地址129.64.134.5的子网掩码，它们的二进制形式表示如下：

IP地址	129.64.134.5	10000001.01000000.10000110.00000101

子网掩码 255.255.128.0 11111111.11111111.10000000.00000000

如果忽略子网掩码，在网络分类的基础上解释该地址129.64.134.5，那么129.64.0.0是网络号，0.0.134.5是主机号，该地址属于B类。

如果使用子网掩码，那么255.255.128.0中连续17个二进制1串，应用到IP地址上，将把它分为以下两个部分：

		网络号	主机号
IP地址	129.64.134.5	10000001.01000000.1	0000110.00000101
子网掩码	255.255.128.0	11111111.11111111.1	0000000.00000000

用十进制表示法，网络号和主机号用0补足到32位的值分别为129.64.128.0和0.0.6.5。

子网掩码的运用可以被解释为执行逻辑与AND操作。

举个例子，IP地址129.64.134.5的网络号就是其与子网掩码255.255.128.0的逻辑与的结果：

(10000001 01000000 10000110 00000101) AND (11111111.11111111.10000000.00000000)

说明 还有其他格式也可以用于子网掩码。如用16进制可以方便解释掩码值：FF.FF.00.00是B类网络的掩码。另外一种经常碰到的表示方式：185.23.44.206/16意味着该地址的掩码包含16个1——即16位为分配给网络号。

掩码机制广泛应用在IP路由中。在路由中，掩码有很多用途。比如，网络管理员可以用它们将ISP（因特网服务提供商）分配给该公司的某类网络分割成多个子网而不用再向ISP申请新的网络号。

这种操作称为**子网化**（subnetting）。基于同样的机制，ISP们可以通过引入所谓的前缀来减少路由表的大小，加入多个网络的地址空间，提高网络路由的性能。这个操作叫做**超网化**（supernetting）。我们将在本章后面介绍CIDR技术时详细讨论它。

17.4 IP地址分配顺序

根据定义，IP寻址方式必须保证网络编号的唯一性以及网中主机编号的唯一性。因此，分配网络号和主机号必须**集中**（centralized）控制，IP地址分配的建议顺序在RFC 2050中描述。

17.4.1 自治网络中的地址分配

对于因特网中的网络而言，编号的唯一性只能通过专为此设立的集中控制机构来协调。单一的、自治的IP网络中网络号和主机号的唯一性可以由网络管理员实现。

这种情况下，网络管理员必须配置整个地址空间，因为IP地址到未连接网络的映射并不会产生负面效果。因为完全足够保证分配的地址的语法正确并且满足前面列出的条件限制（网络号和主机号不能由全0或全1组成），所以管理员可以任意地选择地址。同时，TCP/IP中的主机号与它的本地地址无关。

然而，若使用这种方法，将来无法将该网络连入到因特网上。任意选择的地址可能与集中分配的因特网地址相同，为了避免这种相同引起的冲突，因特网标准定义了几个**私有地址**以供自治网络使用。

- A类网中——网络号10.0.0.0
- B类网中——16个网络号的范围：172.16.0.0——172.31.0.0
- C类网中——255个网络号的范围：192.168.0.0——192.168.255.0

这些地址被集中分配地址排除在外，它们组成一个很大的地址空间，足以为任意大小的网络中的各主机编号。自治网络可以使用以上范围内的地址。注意，不同自治网中的私有地址可以一

致，同时使用私有地址使得将它们正确无误地连入因特网。为这种目的实现的特殊技术[⊖]大大减少了地址冲突。

17.4.2 集中式的地址分配

与因特网相似的大规模网络中，网络地址的唯一性由一个集中式、层次结构的地址分配系统来保证。网络号只能根据特定因特网权威机构的建议来指派。自1998年开始，主要负责注册全球因特网地址的权威机构是**因特网名称和数字分配机构**（Internet corporation for assigned names and numbers，ICANN），这是一个非商业、非政府机构、由一群负责人管理的组织。这个组织控制着各区域部门的操作，范围覆盖各大洲：ARIN（美洲）、RIPE（欧洲）以及APNIC（亚洲-太平洋）。区域部门为大型ISP们分配地址区域，后者则将它们分配给客户，其中可能有一些是小型ISP。

IP地址短缺是集中式分配所面临的问题。很长一段时间里，获取B类地址变得相当困难，事实上申请到A类地址已经变成一种不可能。这个缺点的形成不仅仅是因为网络规模的持续增长，也因为可用地址空间的局限性。通常，C类网络的拥有者可以使用254个地址的一小部分。比如，考虑当有需求通过WAN链路连接两个网络时，我们使用两个根据“点对点”设计相连的路由器（图17-2）。对于一个两个相邻路由器端口相连而组成的退化网络来说，有必要分配一个独立的网络号，尽管这种网络中只有2个节点。

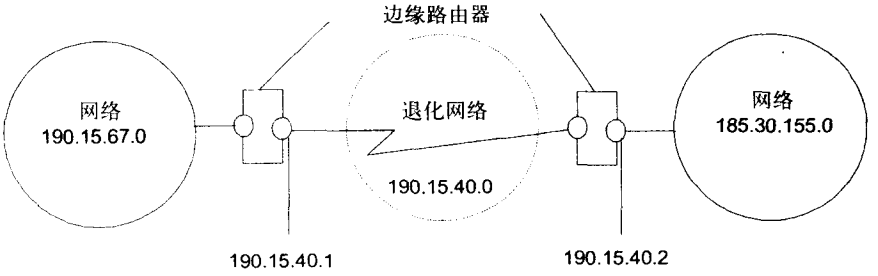


图17-2　IP地址空间利用不足

为了减缓地址短缺的问题，TCP/IP栈的开发者推荐了多种解决方法。主要的创新和有效的方法是移植到新的IP版本IPv6上，这样通过使用16字节地址大大增加了可用地址空间。然而，即使是当前的IP版本，IPv4，也支持一些保证更高IP地址利用效率的技术，如NAT和CIDR.

17.4.3 寻址和CIDR

CIDR于1993年正式引入并且被标准化在RFC 1517、RFC 1518、RFC 1519和RFC 1520中，它允许地址分配中心为它们的用户分配所需的某个地址号。

在CIDR技术中，IP地址根据长度可变的掩码而不是预先指定的位数来划分网络号和主机号。该掩码由服务提供商指定给用户。为了应用CIDR，管理地址的机构必须拥有连续的地址空间，这类地址有着相同的**前缀**（prefix）（即多个最高有效位的值相同）。假设某服务提供商有2^n个连续的IP地址，这样，前缀的长度为（$32-n$）位，余下的n位扮演着序列号计时器的角色。

当有客户端向服务提供商申请特定范围的地址时，服务提供商根据需求的地址数目将连续的地址空间“砍成”三段：S1、S2、S3，同时必须满足下列需求：

- 分配区域的地址数必须为2的指数

⊖ 第20章中介绍的网络地址转换技术就是这个技术的一个例子。

• 分配地址池的初始边界必须为所需主机数的倍数

图17-3中显示的每个区域的前缀都有它自己的长度，给定区域内的地址数越少，前缀就越长。

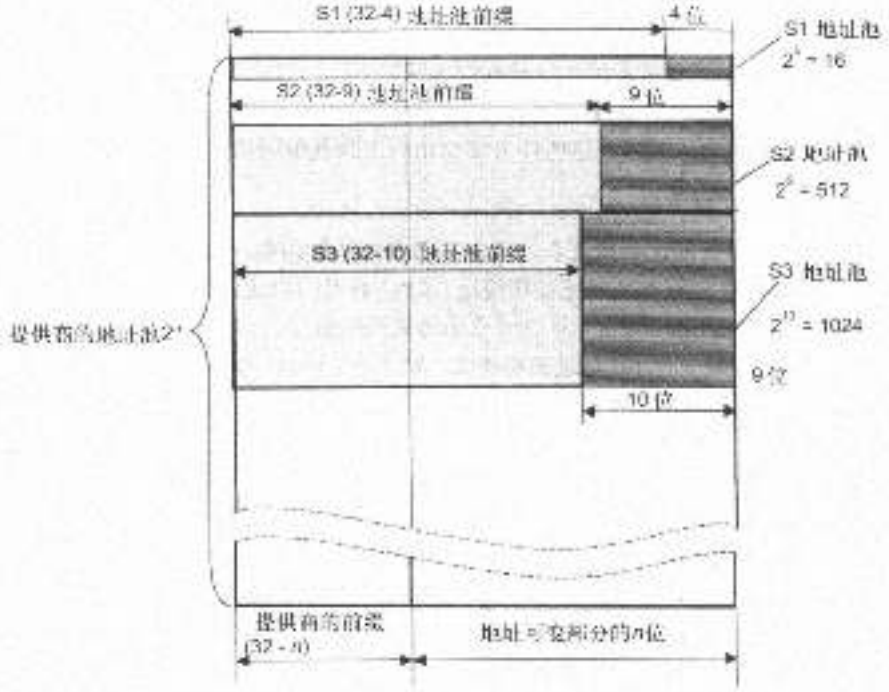


图17-3 基于CIDR技术的地址分配

示例 假定ISP有一个范围从193.20.0.0到193.23.255.255（1100 0001.0001 0100.0000 0000.0000 0000－1100 0001.0001 0111.1111 1111.1111 1111）的地址池，这味着该ISP可分配的地址数为2^{18}。因此，ISP的前缀有14位——1100 0001.0001 01这者是一种表达方式，193.20/14。

如果该ISP的某个用户需要少量的地址——比如，13个——那么ISP可以提供好几种选择：193.20.30.0/28子网，193.20.30.16/28子网和193.21.204.48/28子网。任何一种选择，该用户都有仅四位可供配置主机号，因此我们用满足用户需求（13）的最小数来表示分配给用户的节点数，用2的指数形式表达为2^4 = 16。这种情况下每个分配池的前缀减4作网络号，共有32－4＝28位。

现在，考虑大型公司客户向ISP申请服务的另外一个情况。这个客户可能计划自己提供国特网接入服务。假定该需要4 000个主机的地址块，为这样的大型网络编号，至少需要12位。这味着分配的地址池的大小比需求的数目略大一些，算下来应该是4 096个。分配池起始的边界必须是分配池大小的倍数，满足要求的地址有：193.20.0.0, 193.20.26.0, 193.20.32.0, 193.20.48.0或者其他以12个0结尾的地址。假定ISP向该客户提供了193.20.16.0到193.20.31.255的地址范围，对该范围而言，聚合网络号（前缀）有20位，为193.20.16.0/20。

由于CIDR，ISP承担起了根据每个客户的需要和需求将地址空间分块的职责。

在第18章中，我们会重提CIDR技术，解释它是怎样有效地节约地分配地址同时提高路由效率的。

17.5 将IP地址映射到本地地址

创建IP时必须解决的一个主要问题就是由多个子网组成的互联网之间的协调操作，通常要用

到不同的网络技术。IP分组跨互联网传递时，经常发生组成网络中TCP/IP与本地技术的相互干涉。在每个路由器上，IP决定该分组将被传递到本网中的下一个路由器的地址。为了解决这个问题，本地协议决定下一个路由网络接口的IP地址（如果当前已经是目的网络的话，就是主机地址）。为了使本地网络技术能够将分组传送到下一个路由，有必要执行下列操作：

1）将分组封装到与该网络相适应格式的帧中（比如以太网）

2）将该帧及其本地地址提供给下一个路由

正如已经说明的那样（见第4章4.4.4节），这些任务都交由TCP/IP栈的网络接口层来完成。

17.5.1 ARP

既然本地地址和IP地址之间没有任何依赖关系，建立两者之间映射的唯一方法就是用对照表。经由网络配置后，每个接口都知道它的本地地址和IP地址。该映射可以看成是一张分布在各个网络接口上的表格，唯一的问题在于如何在网络主机之间交换该表格的信息。

为了根据IP地址来定义本地地址，设计了**地址解析协议**（Address resolution protocol，ARP）。根据本地网络上数据链路层协议的不同，ARP的实现也有所不同。你应该能想到，这可能是LAN协议（以太网、令牌环或FDDI）的一种，可以同时广播访问所有的网络节点；或是WAN协议（X.25或帧延迟）的一种，按照规则不支持广播访问。

现在来考虑一下支持**广播**（broadcasting）的LAN中的ARP操作。

图17-4显示了包括两个网络的IP网络中的碎片情况：以太网1（包含三个终端节点：A、B和C）和以太网2（包含D、E两个终端节点），这些网络分别连在路由器接口1和2上，每个网络接口都有一个IP地址和一个MAC地址。假定某些情况下，C主机的IP模块往D主机发送分组，C节点的协议已经判断出下一个路由器的IP地址，IP1。在将数据封装成以太网帧发往下一个路由器之前，需要确认一下其相应的**MAC地址**（MAC address）。IP向ARP发出请求以解决这个问题。ARP在每个网络适配器或路由器的接口上维护着一张独立的**ARP表**（ARP table），在网络运行的过程中，这张表格积累着本网络中其他接口IP地址与MAC地址之间的对应信息。最初当计算机或路由器连入网络时，所有的ARP表格均为空。

- 在图17-4中，步骤（1）对应的是将下列消息从IP传到ARP："拥有IP地址IP1的接口对应的MAC地址是什么？"
- ARP开始查询相应网络接口的ARP表（图17-4中的步骤（2））。假设当前记录中没有被查询的IP地址。
- 将被传送的IP分组无法从ARP表决定本地地址，存于缓存中；ARP产生一个申请，将它封装成以太网帧，并且广播出去（图17-4中的步骤（3））。
- 以太网1网络中的所有接口都收到这个**ARP请求**（ARP request），将它转入"本地"ARP。ARP将分组中指定的IP1地址与它所到达的接口地址相比较，最终找到一个匹配（本例中为路由器1），并且生成一个**ARP应答**（ARP reply）消息（图17-4中的步骤（4））。

在ARP应答中，路由器提供它的本地地址MAC1，并且用它的本地地址将ARP应答传送到发出申请的节点（本例中为节点C）。这种情况下，既然ARP请求的格式提供了发送端的本地和网络地址，所以并不需要广播应答。注意，因为路由器限制广播帧，所以这样的ARP请求传播的领域仅局限在以太网1中。

图17-5显示了封装ARP消息的以太网帧，ARP请求和应答均用该格式，表17-2列出了使用以太网传输实际ARP请求的字段值。

网络类型（*network type*）字段的值为1代表着以太子网。

协议类型（*protocol type*）字段的允许使用ARP解析IP以及其他协议，对于IP而言，这个字段

值为0X0800。

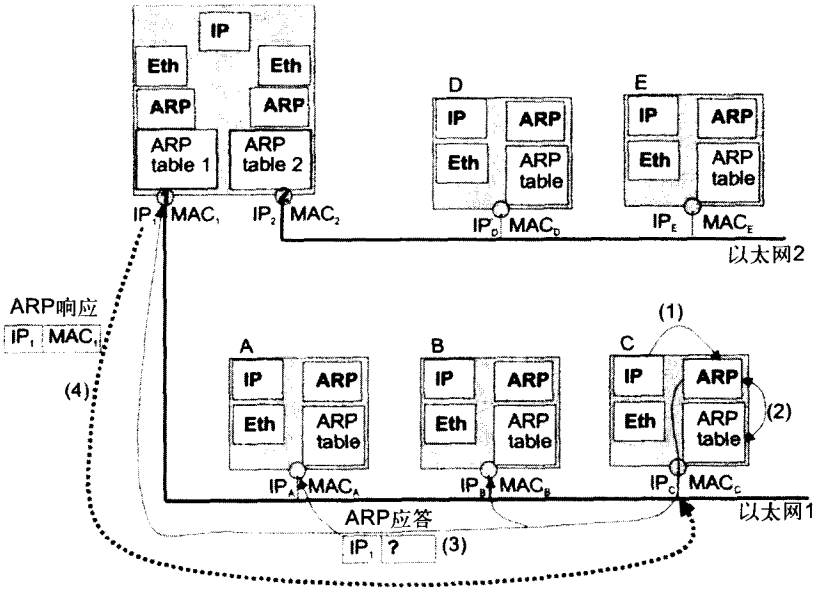


图17-4 ARP操作方法

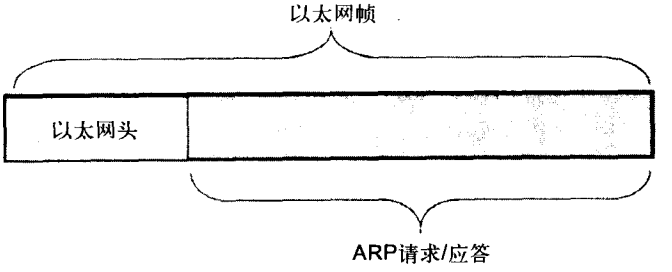


图17-5 在以太网帧中封装ARP消息

以太网协议的本地地址长度为6个字节；IP地址长度为4字节。ARP请求的操作域的值为1代表着请求，而2代表着应答。

从这个请求开始，紧接着在以太网中，IP地址为194.85.135.75的主机试图去判断来自同一个网络中另一个IP地址为194.85.135.65的主机的MAC地址。被请求的本地地址域全填为0（*the field of the requested local address is filled with zeros*）。

表17-2 ARP请求的示例

网络类型	1（0x1）
协议类型	2048（0x800）
本地地址长度	6（0x6）
网络地址长度	4（0x4）
操作	1（0x1）
发送端的本地地址	008048EB7E60
发送端的网络地址	194.85.135.75
接收端的本地地址（被请求的）	000000000000
接收端的网络地址	194.85.135.65

识别出IP地址的主机向这个请求发送一个应答。如果这个网络中没有请求IP地址的主机，则不会产生ARP应答。这种情况下，IP将丢弃发送给它的IP分组，表17-3列出了可能发送给前一个ARP请求的ARP应答的字段值。

通过交换这两个ARP消息，已从接口194.85.135.75发送请求的IP模块判断出与IP地址194.85.135.65相对应的MAC地址是00E0F77F1920。这个地址将被放在封装了IP分组并且准备发送的以太帧的头部。

为了减少网络中ARP消息的个数，识别出的IP地址与MAC地址之间的映射关系保存在适当接口的ARP表中。本例中，记录应该如下所示：

194.85.135.65　　00E0F77F1920

在ARP模块完成对ARP应答分析的若干微秒之后，一条新的记录被自动加入到ARP表中。现在，如果又有分组需要发往194.85.135.65，那么IP在发送广播请求之前，会自动检查该地址是否已存在于ARP表中。

表17-3　ARP应答的示例

网络类型	1（0x1）
协议类型	2048（0x800）
本地地址长度	6（0x6）
网络地址长度	4（0x4）
选择	1（0x1）
发送端的本地地址	00E0F77F1920
发送端的网络地址	194.85.135.65
接收端的本地地址（被请求的）	008048EB7E60
接收端的网络地址	194.85.135.75

ARP表不仅仅通过到达接口的ARP应答来增加条目，还可以从广播ARP请求中检索有用信息。事实上，正如表17-2和表17-3中所示，每个请求都含有发送端的IP地址和MAC地址。所有接收到该请求的接口都可以将本地到网络地址的映射关系存放入自己的ARP表。特殊情况下，表17-2中所有接收到ARP请求的主机可以把下列记录存入它们的ARP表中：

194.85.135.75　　008048EB7E60

因此，增加了上述两条映射记录的ARP表，如表17-4所示。

表17-4　ARP表示例

IP地址	MAC地址	记录类型
194.85.135.65	00E0F77F1920	动态
194.85.135.75	008048EB7E60	动态
194.85.60.21	008048EB7567	静态

记录类型域可取下列两值之一：静态或动态。静态记录使用ARP工具手动生成，不存在过期问题。更精确地说，它们会一直生存到主机或者路由器关机。

动态记录则是ARP模块通过LAN技术的广播功能获得。ARP表不仅通过分析到达当前接口的ARP应答来增加条目，还可以从广播ARP请求中检索有用信息。从表17-2中可以明显看出，ARP请求，除了其他信息外，还包含发送端的IP地址和MAC地址。即使没有与被申请地址的匹配，主机也会将这个有用的信息存入自己的ARP表中。

动态记录必须定期刷新，如果在预设的时间间隔内记录没有更新（几分钟），该记录将从表中删除。因此，ARP表存放在网络中活跃的而不是所有节点上。因为这种存储信息的方法叫做缓存，ARP表有时也被称为**ARP缓存**（ARP cache）。

说明　有些IP和ARP的实现并不将IP分组放入等待ARP应答的时间队列里。取而代之的是，它们只是丢弃IP分组，把恢复的任务交给TCP模块或者使用UDP操作进程的应用程序来代理。这种恢复使用超时重传机制，消息的重传因为第一次尝试已经更新了ARP表而总是成功。

WAN中的地址解析方法完全不同。想想看，WAN不支持广播消息。这种情况下，网络管理员必须手动创建ARP表并将他们放到某个服务器上去。这些ARP表说明，比如说，IP地址到X.25地址的映射，后者被IP解析为本地地址。WAN中，还有一个ARP自动化的趋势，从所有连接到WAN的路由器中选出一个专职路由器，该路由器维护所有其他主机和路由器的ARP表。

当使用这种集中控制方法时，所有的主机和路由器只需要手动指定专职路由器的IP地址和本地地址。每次开启时，主机和路由器都将自己的地址注册到该专职路由器。任何时候需要从IP地址判断本地地址时，ARP模块将向专职路由器递交申请并且自动得到应答，无需管理员的参与。以这样方式操作的ARP路由器被称为ARP服务器（ARP server）。

某些情况下，需要解决一个反向问题，也就是，从已知的本地地址获取IP地址。这种情况下，即可使用反向地址解析协议（*reverse address resolution protocol*）（反向ARP或RARP）。该协议适用于，举个例子，当一个无磁盘工作站启动时，它不知道自己的IP地址而只知道网络适配器的

MAC地址的情况。

17.5.2 代理ARP

代理ARP（Proxy-ARP）是ARP的变体之一，它允许将IP地址映射为支持广播功能的网络中的硬件地址，即使被请求的主机位于当前网络的边界之外。

图17-6显示了只有一个终端节点（计算机D）的网络，该节点操作在远端主机模式。第四部分的第23章将详细讲解该操作模式。目前只需要知道这种模式下的终端节点具备以太网中计算机的一切特性，它有属于同一网络的IP地址IP_D。对于以太网中的所有节点来说，连接远程节点（调制解调器、拨号网络、PPP的存在）的特定特性绝对是透明的，并且它们用正常的方式与该主机进行交互。为了使这样的操作模式成为可能，代理ARP应运而生。因为远程主机通过PPP相连，所以它没有MAC地址。

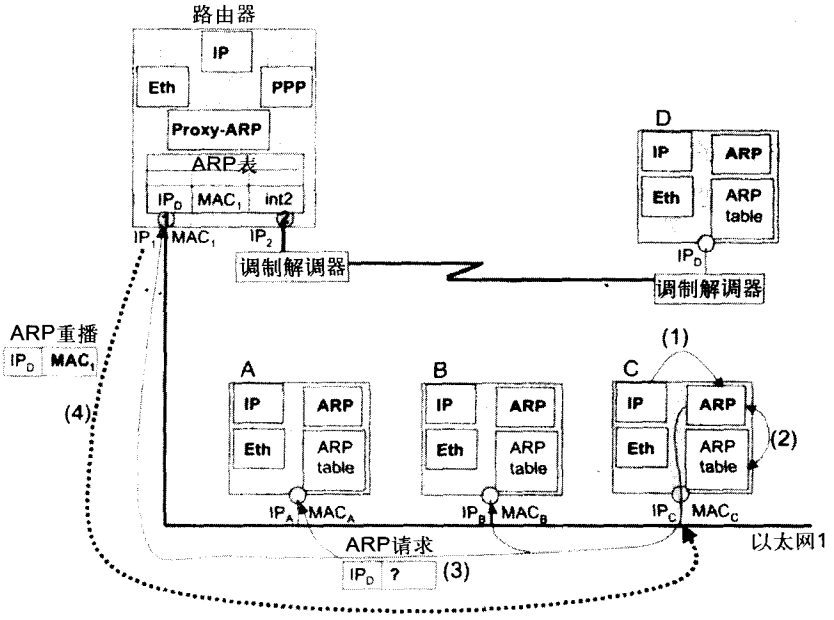


图17-6 代理ARP操作方法

假定计算机C上的应用程序想要往计算机D发送分组，它知道目的IP地址，IP_D；然而，正如前面提到的，要想利用以太网传输分组，需要将它封装成以太网帧从而必须提供MAC地址。为了判定计算机D的MAC地址，计算机C的IP申请ARP，发送出一条包含ARP请求的广播消息。若路由器上没有安装代理ARP，那么将不会有主机响应这条请求。

然而，代理ARP安装了并且如下操作：当远程主机D连入网络时，路由器的ARP表中新增如下记录：

IP_D——MA_{C1}——int2

该记录的意义如下：

- 当与IP_D地址关联的ARP请求到达时，ARP应答必须携带路由器接口1的硬件地址MAC1。
- 具有IP_D地址的主机连接到路由器的第2个接口。

因此，安装了代理ARP的路由器将响应主机C发送的广播请求。路由器将它自己的硬件地址MAC1取代计算机D的MAC地址填入“代理”ARP应答中。

主机C，没有察觉到任何“小伎俩”，向MAC1地址发送封装了IP分组的帧。收到该帧之后，

代理ARP“意识到”这个帧并不是给它的，因为分组含有另外一个主机的IP地址。因此，必须在ARP表中查询目的地址，查询结果显示该帧必须传递到连接在第二个接口上的主机。

这是代理ARP最简单的应用；然而，充分理解后，它体现了该协议的操作逻辑。

17.6 DNS

17.6.1 平面符号名称

在最初为LAN开发的操作系统中，如Novell NetWare、Microsoft Windows或IBM OS/2，用户总是使用计算机的符号名称。因为LAN由少量计算机组成，所以人们采用了*平面名称*（*flat names*），即一串无分割的文本字符串，如NW1_1、mail2、LONDON_SALES_2等。为了建立符号名称到MAC地址的映射，这些操作系统采用广播请求机制，与ARP使用的类似。打个比方说，NetBIOS协议，作为很多LAN操作系统的基础，实施了广播名称解析机制。所谓的NetBIOS名称作为LAN中平面名称的主要类型存在了很多年。

对于TCP/IP栈而言，通常服务于大型、地理分布广泛的网络，这样的解决方法显然是远远不够的。

17.6.2 层次式符号名称

TCP/IP栈采用的是域名系统DNS，一个允许在名字中使用任意组合的层次结构（图17-7）。

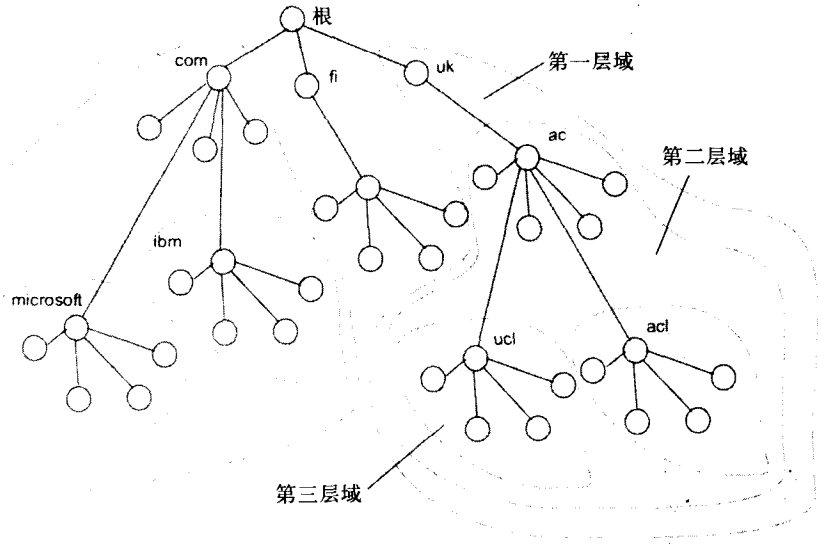


图17-7　域名空间

域名的层次与非常流行的文件系统中所采用的文件名的层次相类似，名称树始于树根，这里用点“.”表示。紧随树根的是名字中最重要的符号部分，然后是次重要的符号，以此类推。名字中最不重要的部分与网络终端节点相对应。文件名领先的是高位的组成部分，随后是稍低层的组成部分，如此继续下去。与文件名不同的是，域的名称始于低层的组成部分，止于最高层的组成部分。域名的各个组成部分之间用点相隔，如partnering.microsoft.com，其中partnering是microsoft.com域中的一台计算机。

将名称分拆同时也拆开了管理员指派唯一名字的责任并且在层次的特定限制内将他们分成多个人或组织。因此，对于图17-7中的示例，才有可能将确保所有以“us”结尾的名称给予更低层次的组成部分唯一不冲突的名字的责任施加给某个个体。如果该个体能够承担起这份责任，那么所有类

似www.us，mail.mmt.us或者m2.zil.mmt.us的名字根据它们的重要程度在第二个字段上加以区分。

管理责任的分离解决了产生唯一名字的难题，而不需要花费大量精力协调负责同一层次的多个组织之间的名字问题。很显然，必须有一个负责顶层层次名字分配的组织存在。

包括一个或多个组成部分的最重要部分的名称集合组成一个域。打个比方，www1.zil.mmt.ru、ftp.zil.mmt.ru、yandex.ru以及s1.mgu.ru都属于域ru，因为他们有着共同的名称的最重要部分——ru。另外一个示例是mgu.ru域，在图17-8的名字中，下列属于该域：s1.mgu.ru、s2.mgu.ru以及m.mgu.ru，这个域的两个重要组成成分总是mgu.ru的名称组成。mgu.ru域的管理员负责下一个层次名字的唯一性，比如s1、s2和m，所生成的子域s1.mgu.ru、s2.mgu.ru和m.mgu.ru都是mgu.ru的子域。为了简洁起见，子域通常由最不重要的组成成分命名——s1、s2以及m。

注释 术语“域”有很多含义；因此必须与特定的上下文一起解释。除了TCP/IP栈的域名以外，计算机词库中还有Windows NT域、冲突域等等。这些域概念的一致性在于都指的是具备特定共同特性的计算机集合。

如果每个域或子域中下一层次名字的唯一性都得到了保证，那么整个名字系统中域名的唯一性也得到了保证。

与文件系统相仿，DNS也提供*短名称*、*相对名称*以及*完全规范域名*（*FQDN*）。短名称是网络终端节点的名字，比如主机或者路由端口。短名称代表着域名树上的叶子节点。相对名称是起始于某一层次的复合名字（然而，不是从顶上开始）。比如，www1.zil是一个相对名称。而FQDN则包括所有层次的名字成分，从短名称开始直至树根：www1.zil.mmt.ru。

根域由集中式因特网权威机构：IANA和InterNIC统一管理。他们在集中控制的基础上负责将顶层域分配给各个国家。这些域名必须遵从ISO 3166国际标准，国家代码为指定国家的2～3个字母的缩写，如us（美国）、ru（俄罗斯）、uk（英国）以及fi（芬兰）。对于不同种类的组织，存在着下列缩写：

- com——商业组织（如microsoft.com）
- edu——教育组织（如mit.edu）
- gov——政府组织（如nsf.gov）
- org——非商业组织（如fidonet.org）
- net——维护网络的组织（如nsf.net）

每个域都由一个独立的组织管理，他们通常把域分成子域并且把这些子域的管理权限委托给其他组织。为了获取域名，需要向一个InterNIC已经授权具有分配域名权力的特殊组织进行登记。

要点 计算机根据它们的FQDN被划分到各个域，同时，他们拥有属于不同网络和子网的独立IP地址。比如，mgu.ru域可能包含地址为132.13.34.15、201.22.100.33以及14.0.0.6的主机。

DNS在因特网上执行，然而，它也可以作为自治域名系统在使用TCP/IP栈并且未连到因特网的企业网运行中。

17.6.3 DNS的操作方式

与ARP类似，建立符号名称和本地地址映射的广播方法，只在未被分割成子网的小型LAN中有效。在大型网络中，不支持无限制广播，需要用到另外一种符号名称解析的方法。支持网络中所有计算机不同类型地址的映射的集中式服务是广播方式很好的替代方法。比如，微软在Windows NT中实施了集中式WINS服务，WINS服务支持NetBIOS数据库以及与之对应的IP地址。

TCP/IP网络中域名和IP地址的映射关系可以通过本地主机工具或者集中式服务来建立。

在因特网革命的早期阶段，必须在每个主机上手动创建一个名为hosts.txt的文本文件，此文件

包含一定数目的记录，每个记录为“IP地址——域名”的关系对，比如，102.54.94.97——rhino.acme.com。

随着因特网的发展，主机文件不断增大。因此，必须开发一个可延拓的名称解析的方法。DNS就是解决方案之一。

DNS是一种基于域名与IP地址之间映射关系的分布式数据库的集中式服务。在其操作过程中，DNS使用客户端-服务器协议，此协议定义了DNS服务器和DNS客户端。DNS服务器支持映射的分布式数据库，DNS客户端请求服务器完成从域名到IP地址的解析。

DNS服务使用与主机文件相类似的文本文件。这些文件由网络管理员手工创建。然而，DNS服务依赖于域的层次结构，每个DNS服务的服务器只能存储网络名称的部分内容而不是像主机文件那样存储全名，当网络中的主机数目不断增大时，规模问题可以通过创建新域和子域以及增加新的DNS服务器来解决。

必须为名字中的每个域创建一个DNS服务器，服务器上名字的分布有两种方法。第一种情况下，服务器可以存储整个域中“域名——IP地址”的映射，包括所有的子域。然而，这种解决方法可延拓性很差，因为随着新的子域不断添加，服务器上的负载可能会超出它的承受能力。因此，另外一种解决方法更加常用，这种方法中，域名服务器只存储在低一级层次的部分名字，而不是域名。这个方法有点儿像文件系统目录的使用，只包含该目录中文件以及子目录中文件的记录。使用这种方法组织DNS服务，负载会平均地分摊到网络中所有DNS服务器上。比如说，第一种情况下，域mmt.ru的DNS服务器会存储所有以mmt.ru结尾名字的映射：www1.zil.mmt.ru、ftp.zil.mmt.ru、mail.mmt.ru等等。而第二种情况下，该服务器只会存储类似于mail.mmt.ru、www.mmt.ru等名字的映射，所有其他的映射关系必须存储在zil子域的DNS服务器上。

除了映射表以外，DNS服务器还包含到其子域的DNS服务器的链接，这些链接将孤立的DNS服务器连成了统一的DNS服务。引用方法是相应服务器的IP地址。多个可选的DNS服务器专职服务于根域，它们的IP地址被广泛公开（比如，可以通过InterNIC来获取它们的列表）。

假设已经得知符号名称，解析DNS名字的步骤在很多方面与在文件系统中寻找文件地址的步骤相似。两种情况下，完全规范名称都反应了某引用表的层次组织结构——分别为文件目录和DNS表。这里的域和它的DNS服务器可类比于文件系统。与符号文件名称相似，域名也具备名字独立于其物理位置的特征。

通过符号名称查找文件地址的步骤包括从根目录开始顺序搜索所有的目录。这时，首先检查缓存和当前目录。对于根据域名查找IP地址而言，也需要从根域开始搜索所有形成子域链的DNS服务器，直到找到主机。文件系统搜索和DNS搜索最大的区别在于文件系统位于单个计算机中，而DNS是分布式的。

解析DNS名称有两种方法，第一种情况下，DNS客户端协与调搜索IP地址相关的工作。

- DNS客户端向根DNS服务器请求，申明FQDN。
- DNS服务器做出响应，在被请求名称的最高有效部分中指出服务于上层域的下一个DNS服务器的地址。
- DNS客户端向下一个DNS服务器提出申请，该DNS服务器将请求重定位到所需子域的DNS服务器上，如此类推。本步骤一直持续到所需名称到特定地址的映射关系找到为止，服务器将最终结果返回给客户端。

这个交互的方法被称为**非递归**（nonrecursive）或**迭代方法**（iterative method）。这里，客户端向不同域名服务器递交一系列请求，因为客户端需要执行的任务相当复杂，本方法很少使用。

第二种方法使用**递归过程**（recursive procedure）：

- DNS向本地DNS服务器（比如说，为客户端名称所属的子域服务的那个DNS服务器）提出请求。
- 如果本地DNS服务器知道答案，立刻返回给客户端。这种情况可能发生在请求的名称与客户端处在同一个子域中；也可能发生在该服务器为另外一个客户端刚刚查询过并且结果依然存储在缓存中。
- 若本地服务器不知道答案，它将如同客户端在第一条中所做的那样向根服务器迭代查询。一旦接收到答案，DNS服务器立即发送给客户端，后者在请求执行期间一直保持着等待本地DNS服务器反馈的状态中。

这种方法中，客户端将查询过程委托给服务器。因此，这种方法被称为非直接或者递归的。实际应用中所有的DNS客户端都采用递归的方法。

为了加速查询IP地址的过程，DNS服务器为经过它们的所有响应广泛采用缓存机制，为了使DNS服务及时地处理网络中发生的变化，响应只在相当短的时间内被缓存，通常为几个小时到几天。

17.6.4 反向搜索区域

DNS服务目的不仅仅在于通过主机名找到IP地址，也在于解决**反向问题**（inverse problem），即，通过已知的IP地址找到主机名称。

当用户只申明了地址或者当从程序从网络上收到的分组中只抽取出地址时，多数程序和应用使用DNS尝试来根据该地址发现主机名。即使对于有着直接记录的地址，反向记录也并不需要存在。管理员或许只是忘记创建它们。有时，创建这样的记录需要额外的代价，特别当反向搜索区域的主服务器由ISP支持时。在反向搜索区域不存在时，因为不得不花大段时间等待反向请求，这种程序操作会有相当大的延迟。

因此，因特网上通过组织所谓的**反向搜索区域**（reverse lookup zone）来解决反向问题。

反向搜索区域（reverse lookup zone）指的是存储着某网络的IP地址和同一网络中主机名称映射关系的表的集合。为了组织一个分布式服务并且使用同一软件来搜索名称和地址，我们考虑了与复合名称风格相似的复合IP地址。

比如说，类似与192.31.106.0的地址被分为包含最高位与域194对应的地址，紧随其后的是包含域106的域31。为了存储所有以192打头的地址，194范围以及它自己的域名服务器——主用和备用的——被创建了。在写入一个地址时，地址的最高位部分是最左边的；而对于名字，情况正好相反。因此，为了在反向请求中实现完全对应，地址以反向顺序指定（如，上例中为106.31.192）。

对于服务器上控制着根层次反向搜索区域的记录，一个特殊的范围——in-addr.arpa——被创建了。因此，示例中地址的完整记录如下所示：

106.31.192.in-addr.arpa.

反向搜索区域的主要服务器使用独立于主区域文件的数据库文件，它有着相同名称和地址的直接映射记录。这样的组织会导致不一致性，因为同样的映射关系文件中存储了两次。

17.7 DHCP

对于网络中的正常操作而言，计算机或路由中发送或接收IP分组的每个网络接口都必须指定IP地址。

IP地址的指派必须在配置接口过程中手动执行。对于计算机而言，这个过程包括填充屏幕上显示的一系列对话框。当用这种方式处理时，管理员必须牢记可用的地址集中哪些地址已经被分配出去，那些还没有分配。当执行配置步骤时，管理员必须为客户端指派一系列参数，包括IP地址以及其他TCP/IP有效操作时所需的参数，如子网掩码、网关地址、DNS服务器地址以及计算机

的域名。即使对小型网络而言，这也是一个例行有时甚至乏味的操作。

动态主机配置协议（dynamic host configuration protocol，DHCP）自动进行配置网络接口的过程，通过使用集中管理的地址数据库来确保消除重复地址的出现。DHCP操作在RFC 2131和RFC 2132中描述。

17.7.1 DHCP方式

DHCP根据客户端-服务器模式来运作。系统启动时，DHCP客户端向网络发送一个广播请求，申请分配一个IP地址。DHCP服务器响应请求并且发送一条包含IP地址和其他配置参数的应答信息。

DHCP服务器可以在多种模式下操作：

- 手动分配静态地址
- 自动分配静态地址
- 自动分配动态地址

在所有的操作模式中，管理员需要为DHCP服务器指定一个或多个范围的可用IP地址，所有这些地址必须属于同一个网络（即，他们的网络号必须相同）。

在*手动模式*（*manual mode*）中，管理员除了提供可用地址池以外，还必须向DHCP服务器提供严格定义的IP地址到物理地址或其他客户节点标识的映射信息。使用该信息，DHCP服务器通常会为DHCP客户端指派同一个预设好的IP地址以及其他配置参数[⊖]。

在*自动分配静态地址*（*automatic assignment of static address*）模式中，DHCP服务器为客户端任意择IP地址而无需管理员的干涉，该地址选自可用的IP地址池。选择是基于一个常量进行的，这意味着客户端IP地址和它的身份信息之间存在着固定映射关系，在这一点上和手动分配很相似。此映射在DHCP服务器第一次为客户端分配IP地址时设置，以后所有的IP地址客户端请求都将返回同一个IP地址。

当使用*动态地址分配*（*dynamic address distribution*）时，DHCP服务器为客户端分配一定时间内有效的IP地址，也就是**租用期限**（lease duration）。如果某个DHCP客户端离开了网络，那么分配给它的IP地址将自动释放。当该计算机连接到另一个子网时，它会自动获得新的IP地址。终端用户和网络管理员均不参与这个过程。

自动地址分配为重用被释放地址、将来指派给其他计算机提供了可能，因此，除了DHCP可以自动完成管理员配置TCP/IP栈例行工作的优点之外，动态地址分配还使得IP网络中节点数可以超出可用的IP地址数。

示例　考虑一下地址池动态分配的优点，假设某公司职员的大多数工作时间是在家里办公或者出差，在办公室办公的时候每个职员都会用笔记本连入公司IP网络。那么试问，这样的公司需要多少IP个地址呢？

第一个答案如下：地址数必须等于网络中工作的员工数。比如，有500个员工必须拥有IP地址和办公场所。因此，管理员必须向ISP申请两个C类网络，然而，既然该公司职员大部分时间都不在办公室，显然如果采取这个方案的话，这些资源多数时间都处于闲置状态。

另外一个方法是申请经常办公室里员工总数的IP地址数目（同时保留一些备用）。比如，如果通常办公室里职员数不超过50，那么申请64个地址并且配置64个计算机连接器的网络就已经足够。然而，这里存在着另外一个问题——谁以及怎样来配置频繁加入和离开网络的计算机呢？

解决这个问题有两个途径。第一个，管理员（或者移动用户）可以在需要连入公司网络时手动配置。这个方法需要大量例行操作；因此不是个好方法。在这种情况下，自动分配

⊖ 为简便起见，我们省略了这个说明。

DHCP地址看起来十分具备诱惑力，如果采用这种方法，管理员可以在配置DHCP服务器过程中指定64个地址的范围，然后每个新加入的移动用户只需要物理连接到网络，DHCP客户端就启动了，DHCP客户端申请需要的配置参数并从DHCP服务器自动接收。因此，为了容纳500个移动用户，大小为64的地址池以及64个办公场所已经足够。

17.7.2 动态地址分配算法

管理员通过指定DHCP服务器配置的两个主要参数控制若网络配置的过程：可供分配的地址池和租用期限。租用期限规定了计算机在向DHCP服务器重新申请地址之前，当前IP地址的可用时间长度。租用期限参数视网络用户的工作模式而定。如果这是一个教育机构的小型网络，无数学生携带着笔记本前来做实验，租用期限或许是该工作所需要的时间。如果这是一个公司网络，大多数员工有规律地前来工作，那么租用时间应该更长一点儿——几天或者几个星期。

考虑到客户端会向它发送广播请求，DHCP服务器必须和客户端同处一个子网。为了减少由于DHCP服务器故障而出现的网络崩溃带来的风险，有时候会在网络中安装一个冗余DHCP服务器。这个可选方案与图17-8中网络1的配置相对应。

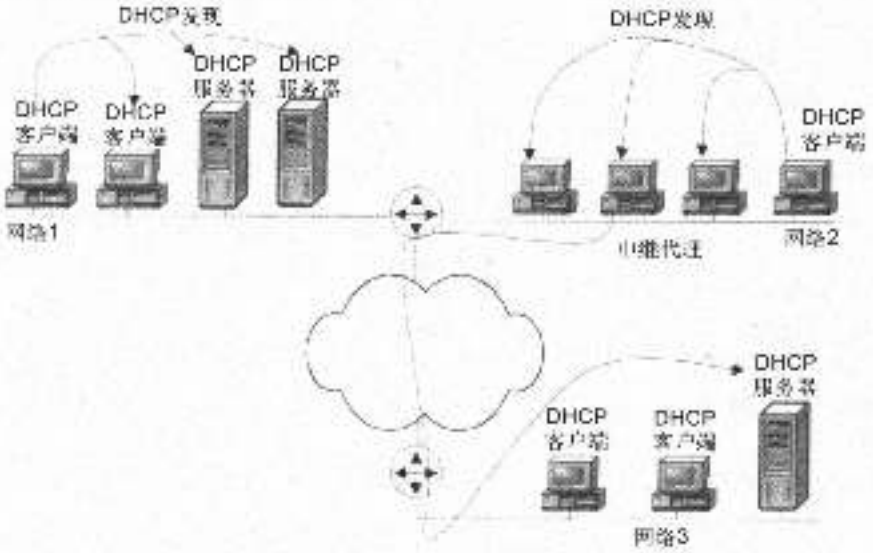


图17-8 DHCP客户端和服务器手动配置方法

有时会出现另外一种情形：网络中没有DHCP服务器。这样，DHCP服务器被DHCP中继代理所取代，后者在DHCP客户端和DHCP服务器之间充当协调器的角色。这种配置的示例可见图17-8中的网络2。代理将本地网络中的客户端请求重定向到位于另外一个子网中的某个DHCP服务器（本例中为网络3）。因此，一个DHCP服务器可以服务于来自多个子网的DHCP客户端。

下一个例子是关于在DHCP服务器和它的服务器部分之间交换信息的简化方法。

当计算机接上电源之后，DHCP客户端发送一条名为*DHCPdiscover*的有限广播消息（目的地址只包含二进制1的IP分组，必须传送到IP网络中的所有主机）。

网络中的DHCP服务器收到这条信息。若该网络不包含DHCP服务器，那么*DHCPdiscover*消息由DHCP中继代理接收。它将这条消息发送给另外一个，可能位于远程网络上——它预先知道IP地址的DHCP服务器。

所有收到*DHCPdiscover*消息的DHCP服务器将它们提供的信息发送到提出请求的DHCP客户端。这些信息放在*DHCPoffer*中发送，每个*DHCPoffer*消息包含一个IP地址和其他配置参数（位于

另外一个子网的DHCP服务器通过代理发送它的响应)。

DHCP客户端从所有DHCP服务器收集配置信息，按照规则，它会选择收到的第一个*DHCPoffer*消息并且发送*DHCPrequest*广播信息，其中包含被接收的DHCP服务器信息（以及配置参数的值)。

所有的DHCP服务器收到*DHCPrequest*消息，被客户端选中的DHCP服务器发送*DHCPacknowledgment*消息（确认IP地址和租用参数)，其他服务器取消它们提供的消息并且将提供的地址归还到可用地址池。

DHCP客户端收到*DHCPacknowledgment*确认之后进入操作状态。

有时，计算机会尝试更新从DHCP服务器获取的租用参数，在租用到期之前它就开始了第一次尝试，向它收到当前参数的同一个服务器递交请求。如果得不到响应，或者响应被拒绝，它会隔段时间重复尝试。假如经过若干次尝试之后，客户端仍然没有从该服务器收到参数，那么它会向其他服务器递交请求。最后，如果新的尝试同样失败，这个客户端就失去了它的配置参数并且进入自动操作模式。

DHCP客户端也可以通过执行DHCP释放命令提前释放租借给它的参数。

在一个IP地址自动分配的网络中，很难判断哪个地址被分配给了特定的主机。这种IP地址的不稳定性带来了一些麻烦。首先，在将符号域名转换成*IP*地址时会遇到困难（there might arise some difficulties when translating a symbolic domain name to an IP address)。试想一下，支持域名到IP地址的映射表的DNS的操作，其中IP地址每两个小时都改变一次！考虑到这个情况，建议为用户频繁访问符号名称的服务器分配静态地址，将动态名称仅留给客户端计算机。然而，在某些网络中，服务器数目如此之多以至于手动配置耗费大量劳动。这种情况导致了增强版本DNS的发展，即动态DNS，其中DHCP和DNS服务协调使用地址信息数据库。

其次，如果用动态分配的地址作为标识符的话，为接口执行远程控制和自动监控变得异常困难（rather difficult to carry out remote control and automatic monitoring)（比如，为了累计统计数据)。

最后，为了确保网络安全，多数网络设备会过滤预先设定域值的分组。换句话说，当使用动态分配IP地址时，*IP*地址的分组过滤变得复杂起来（packet filtering by IP addresses gets complicated)。

最后两个问题通过禁止为监控系统和安全系统的接口使用动态地址可以容易地解决。

小结

- TCP/IP栈使用三种类型的地址：本地地址（又名为硬件地址)、IP地址、符号域名。所有这些地址独立地指派给互联网中的主机。
- 4字节长的IP地址包含网络号和主机号，为了区分网络号与主机号的边界，可以用两种方法。第一种是基于地址分类的方法，第二种则是使用子网掩码。
- 地址分类通过地址的多个起始位来定义。A类地址中，1个字节用来存储网络号，剩余的3字节存储主机号。A类地址用于最大型网络。对于小一点儿规模的网络，C类地址最合适。在C类地址中，网络号占用3字节，只剩下1个字节给主机号。B类网络则居于两者中间。
- 为了把IP地址分成网络号和主机号，启用了与地址相关联的子网掩码。子网掩码的二进制表示形式中值为1的那些位必须被解释成当前IP地址的网络号。
- IP地址在互联网中唯一表示主机；因此必须集中分派。
- 若网络小型且自治，那么该网络中IP地址的唯一性可以由网络管理员确保。管理员可以为网络和主机自由地选择任意的IP地址，唯一要求在于所选的地址必须符合正确的语法。然而，对于自治网络，往往更倾向于使用所谓的私有地址。

- 若网络规模很大，如因特网，那么指派IP地址的任务变得过于复杂从而被分为两个步骤。第一步，网络地址被分散开来，这个步骤由特定管理权威机构来规定以确保网络编号的唯一性。网络号收到之后，第二步，才开始为主机分配地址。
- 将IP地址分派给网络主机可以手动或者自动执行。如果手动分配IP地址，那么网络管理员负责维护已分配和仍然可用的网络地址列表，并且手动配置每个网络接口。当IP地址自动分配时，使用DHCP。此时，管理员预先将可用地址的范围指定给DHCP服务器，服务器应网络主机的要求将IP地址自动分配出去。
- ARP执行网络接口从IP地址到硬件地址映射关系的建立。
- 在支持广播以及不支持广播的网络中，使用两种不同的地址解析方式。以太网、令牌环以及FDDI网络中将IP地址转换成MAC地址的ARP操作执行一个广播ARP请求，到达接口的ARP响应存储在每个网络接口创建的表中。在不支持广播地址的网络中，ARP表统一存储在专用的ARP服务器上。
- TCP/IP栈使用符号名称的域名系统，该系统拥有允许使用任何多成员名字的层次树结构。多个高位名字组成成分相一致的名称组成名称域。如果该网络是因特网的一部分，域名将被统一分配；否则，则由本地分配。
- 域名和IP地址之间的对应关系可以用一个本地主机文件来存储或者利用存储“域名——IP地址”映射关系的分布式数据库的集中式DNS服务来实现。

复习题

1. 分配硬件地址和网络地址的过程的区别在哪里？
2. 以下列出的地址中哪个可以用于IP互联网中的本地地址？哪些地址不能用？
   A. 一个6字节MAC地址，如12-B3-3B-51-A2-10
   B. 一个X.25地址，如25012112654987
   C. 一个12字节的IPX地址，如12.34.B4.0A.C5.10.11.32.54.C5.3B.0
   D. ATM网络的一个VPI/VCI地址
3. 以下描述语句哪些总是正确的？
   A. 每个网桥或交换机的每个接口都有MAC地址。
   B. 每个网桥或交换机都有网络地址。
   C. 每个网桥或交换机的每个接口都有网络地址。
   D. 每个路由器都有网络地址。
   E. 路由器的每个接口总是有MAC地址。
   F. 路由器的每个接口都有网络地址。
4. 以下提供的地址中哪个不能用于因特网主机的网络接口IP地址？对于没有语法错误的地址，判定它所属的分类：A、B、C、D或E。

A. 127.0.0.1	E. 10.234.17.25	I. 193.256.1.16
B. 201.13.123.245	F. 154.12.255.255	J. 194.87.45.0
C. 226.4.37.105	G. 13.13.13.13	K. 195.34.116.255
D. 103.24.254.0	H. 204.0.3.1	L. 161.23.45.305

5. 假设子网内某主机的IP地址为198.65.12.67；该子网的掩码为255.255.255.240。请问子网号是多少，该子网可以容纳多少主机？
6. 假设你知道网络中除了一个以外所有计算机的IP地址到域名的映射，而对于那一台计算机只知道域名，根据这些信息，你能明确地定义出它的IP地址吗？

7. 一个计算机上有多少ARP表？路由器上有多少？交换机上呢？
8. 从功能上来说，ARP可以被分为客户端部分和服务器部分。请描述客户端部分和服务器部分的功能。
9. 请问，管理员会往ARP表中加入什么地址？为什么？
10. 什么情况下需要使用代理ARP？
11. 可能从计算机的域名中断定出它们地理位置上分布的距离吗？
12. 有一台计算机的IP地址为204.35.101.24，域名为new.1firm.net，请断定下列域名中，如果有的话，哪个属于另外一台IP为204.35.101.25的计算机：new1.firm.net、new.firm1.net或new.1firm.net。
13. DNS和文件系统有哪些共同特性？
14. 请问服务于被两个路由器分割的网络，多少DHCP服务器才足够？
15. 为了增强可靠性，网络中有两个DHCP服务器。请问管理员怎样将可用地址池分配给它们中的一个：为每个分配地址池中不重复的部分，或者为两个都分配地址池的公共部分？
16. 为什么反向DNS问题，从名字上看来，通过已知IP地址找到主机名，不能适用于解决正向问题的同样方法（也就是，使用根据名字层次组织成树的相同区域和域文件）？

练习题

1. 假定某ISP有一个B类网络地址可供分配。为了它自己所在网络的主机寻址，该ISP使用了254个地址。如果它们的网络对应于C类地址的话，请断定该ISP可服务的最大客户数。该为将其自身网络连入客户网络的ISP路由器设置什么子网掩码？
2. 如果有一个C类网络可供配置，可组织的子网最大理论数是多少？这种情况下子网掩码扮演什么角色？值是多少？

第18章 因特网协议

18.1 引言

本章的重点为RFC 751所定义的因特网协议（IP）。在沿着IP分组路径的每个下一个网络中，该协议调用当前网络采用的传输工具，将分组传送到连入下一个网络的路由器或者直接传送到目标节点。因此，IP最重要的功能是支持与组成网络基础技术之间的连接。IP功能包括提供连入上层和网络层协议的支持，特别是TCP协议，此协议在TCP/IP栈中执行因特网上数据可靠传输的所有相关任务。

IP是一个无连接协议。这意味着它将每个IP分组都看成是与其他IP分组无关的独立数据单元。IP不具备通常为了确保传输数据的认证机制，如果在分组转发过程中出现错误，IP不会启动任何动作来更正这个错误。比如说，当有分组在传输路由器上因为校验和错误而被丢弃时，IP实体不会试图重传丢失的分组。换句话说，IP实施的是尽力服务策略。

本章将详细描述IP的主要功能，即路由。我们将解释路由表的结构，包括带和不带子网掩码的；还会提供使用固定和可变长度掩码、覆盖地址空间以及实施子网化和超网化技术的示例。最后会调查分组分片的能力IP。

在描述IPv6的特性时，集中在提供更好的可延拓性的现代寻址技术。同时还有IP头部的格式，它通过减少路由器的工作量而使网络带宽得到了提高。

18.2 IP分组格式

分组头部的字段数和为之服务的协议功能的复杂度之间存在着直接的联系，头部越简单，相应的协议也就越简单。大多数协议操作与对分组头部字段携带的控制信息的操作相关。通过研究IP分组头部每个字段的值，不仅可以了解到分组的结构，也可以熟识**因特网协议**（internet protocol）的主要功能。

IP分组由头部和数据段组成，头部包含有下列字段（图18-1）：

版本（version）。该字段占用4位以申明IP版本。当前广泛使用的是IPv4，然而，新的版本IPv6正在逐渐被更多的人使用。

头部长度（header length）。IP分组的头部长度字段也占用4位，它用32位字度量头部长度，通常，头部有20个字节长（5个32位字）。然而，如果增加了控制信息，那么通过附加IP可选字段，长度会进一步增加。头部的最大长度为60字节。

服务类型（ToS）（type of service），或者称呼它的新名字——**区别服务字节**（differentiated services byte）（**DS字节**）。该字段占用1个字节，与之对应的两个变量：ToS（以前的名称）和DS字节（新名称）。不管哪种情况，这个字段用来存储反映分组QoS需求的参数。

在ToS中，这个字段被分为两个子字段，起始的3位组成优先权子字段，优先权可取值0（普通分组）到7（控制信息分组）。路由器和计算机会考虑分组的优先权从而优先处理最高优先级的分组。*ToS*字段还包含3位用来断定**路由选择策略**（route selection criterion），有三个选项：低延迟、高可靠性或者高吞吐量。如果*delay*（D）位设为1，那么在传送这个分组时路由应该选择将延迟最小化。吞吐量（T）位会最大化吞吐量，可靠性（R）位会最大化可靠性。剩余2位是保留位，通常设为0。

20世纪90年代末采用的区别服务标准为这个字段赋予了新的名称并且重新定义了各个位的值。

DS字节只使用该字节的高6位，保留低2位。DS字节字段每位的目的将在第20章描述IP网络中确保QoS方法时介绍。



图18-1　IP分组的头部结构

总长度（total length）。这个字段占用2个字节，申明分组总的长度，包括头部和数据段。分组最大长度受限于该字段的长度，即65 535个字节。然而，多数网络并不使用如此长的分组。当在异构网络上传输分组时，其长度根据携带IP分组的下一层网络协议中定义的最大分组长而定。如果这些是以太网帧，最大分组长为1 500字节，因为该分组必须匹配到以太网帧的数据区域。TCP/IP标准确保所有主机有能力接收576字节长的分组（无论它们是否分片）。

身份认证（identification）。这个字段占用2个字节。用来识别源分组分片所创建的分组。同一个分组的各个分片必须具备相同的标识符值。

标志（flags）。标志占用3位，包含与分片相关的属性。*将不分片*（*do not fragment*，*DF*）位设为1提示路由器不要将此分组分片。但若*多片*（*more fragment*，*MF*）位被设为1，就意味着此分组被分片且本片不是最后一片。剩余位属于保留位。

分片偏移量（fragment offset）。该字段占用13位，申明该分组中数据区域相对于分片源分组的数据区域起始点的偏移量。它用于组装或重新组装分组时用，偏移量必须为8的倍数。

剩余生存时间（time to live）（TTL）。该字段占用1字节，用于指明分组在网络上传输的最大时间间隔。TTL以秒为单位，由发送端指定。分组的当前TTL随着它花费在网络传输中途径的每个路由器上的每一秒而递减一。即使路由器处理分组的时间不足一秒，还是必须将TTL计数器减一。因为如今的路由器处理分组极少需要花费多于一秒的时间，TTL可以被解释成分组路过的转接节点数。如果TTL值在到达目的地之前降为0，则丢弃该分组。因此，TTL参数是分组的一种"自我破坏"的计时器。

协议（protocol）。这个字段占用1字节，包含了数据字段中的信息所要传送的高一层协议的标识符。不同协议的标识符在"分配号码"RFC中列出，RFC 1340一直用到1992年，后来在2002年更新为RFC1700和RFC3232。如今，RFC在定期更新的http://www.iana.org上可以获取。举个例子，序号6表明该分组包含TCP消息，17代表着UDP，1代表ICMP。

头部校验和（header checksum）。这个字段占用2字节，只计算头部的校验和。因为有些头部字段在网络分组传输过程中也会改变（如TTL），必须检查校验和并在每个路由器和终端节点上重新计算。校验和——16位——计算头部中所有16位字和的补码。在计算校验和时，头部校验和字段被设为零。如果校验和不正确，将会报告错误，那么该分组在错误被发现的同时被丢弃。

源IP地址（Source IP Address）和**目的IP地址**（Destination IP Address）。这两个字段长度

相同——32位。

IP选项（IP Options）。这个字段是可选的。原则上是在网络纠错时才使用。该字段包含好多个子字段，每个具有八个预设类型。在这些子字段中，可以指定途径路由器的准确路由、注册分组路过的路由器、存储安全系统数据或保存时间戳。

填充（padding）。既然*IP*选项字段的子字段数是任意的，但有必要在分组头部的最后增加几个字节以将分组对齐到32位边界。这个字段都用0来填充。

下面是一个用微软网络监控（NM）协议分析器从以太网上截获的真实IP分组的头部字段列表。在该列表中，NM提供了字段的十六进制值（括号中）。另外，这个程序有时候用更加用户友好格式的信息来替代字段的数字代码。比如，NM接口用协议名称取代协议字段的代码（本例中它用字符串TCP取代代码6——请看下列粗体的数据）。

IP: 版本 = 4 (0x4)
IP: 头部长度 = 20 (0x14)
IP: 服务类型 = 0 (0x0)
IP: 优先权 = 常规
IP: ···0··· = 普通延迟
IP: ····0···· = 普通吞吐量
IP: ·····0··· = 普通可靠性
IP: 总长度 = 54 (0x36)
IP: 标识符 = 31746 (0x7C02)
IP: 标志摘要 = 2 (0x2)
IP: ....0 = 数据报中的最后一个分片
IP: ......1. = 无法分片的分组
IP: 分片偏移量 = 0 (0x0) 字节
IP: 生存时间 = 128 (0x80)
IP: **协议 = TCP——传输控制**
IP: 校验和 = 0xEB86
IP: 源地址 = 194.85.135.75
IP: 目的地址 = 194.85.135.66
IP: 数据：剩余数据字节数 = 34 (0x0022)

18.3 IP路由方法

考虑图18-2中示例的网间IP路由方法。这个网络将18个网络连入互联网的20只路由器：N1，N2，...，N18代表18个网络号。每个路由器以及*A*、*B*终端节点上都装有IP。

路由器有多个连接网络的接口（端口），每个路由器接口可以看做是一个独立的网络节点：它有一个独立的网络地址和与它相连的网络的本地地址。比如，路由器1有三个N1、N2和N3网络连接的接口。图中，这些端口的网络地址指定为IP_{11}、IP_{12}以及IP_{13}。IP_{11}接口是网络N1的节点，因此地址IP_{11}的网络号字段包括数字N1。同样，地址IP_{12}是网络N2的节点，地址IP_{13}代表网络N3节点。因此，路由器可以被认为是一组节点，每个节点都是单个网络的一部分，路由器作为统一设备既没有网络地址也没有本地地址。

说明 如果路由器有控制单元（如SNMP控制单元），那么它就有本地地址和网络地址，以供中心控制站访问。

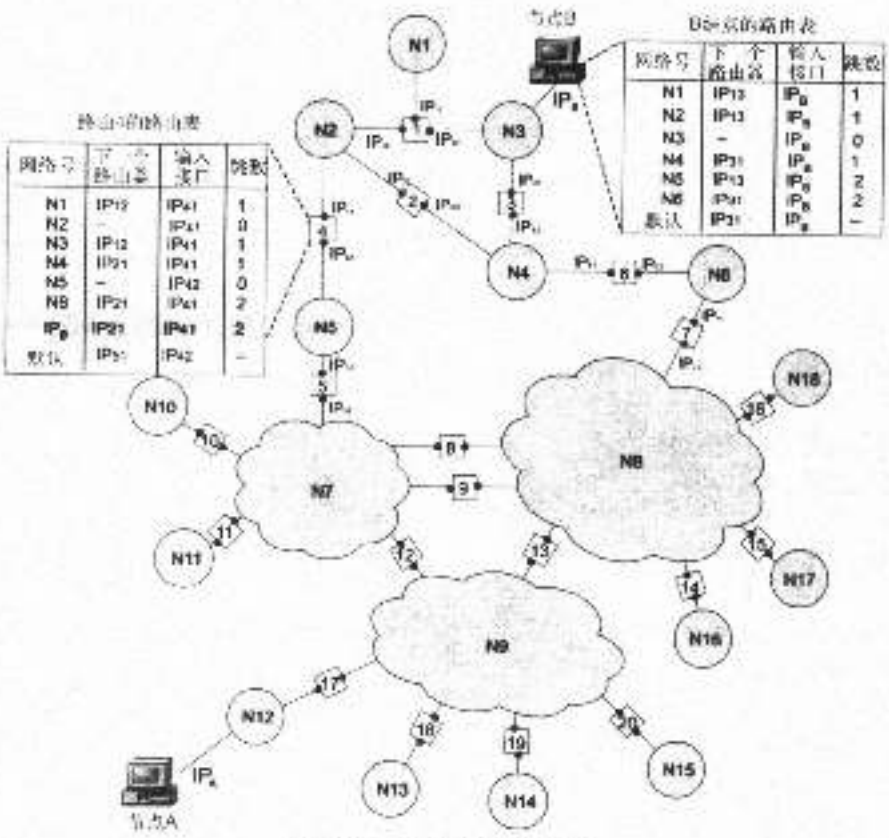


图18-2　互联网中的路由原理

在复杂的互联网环境下，两个终端节点之间的分组传输通常有多个路由。比如说，从节点A到节点B的分组可以通过路由器17、12、5、4、1或者通过路由器17、13、7、6、3，而且在节点A和B之间再多找几个路由也不是难事。

选择路由的问题由路由器和终端节点来解决，根据提供给这些设备的当前网络配置信息以及路由选择规则来进行选择。通常，单个分组的路由传输延迟、分组序列的平均路由带宽或者甚至仅仅考虑到沿着路由传输的中继路由数目的简单策略都是路由选择的评判标准，关于路由信息的分析结果最后存入**路由表**（routing table）中。

18.3.1　简化的路由表结构

我们使用传统表示法来表示图18-2中所示路由器的网络地址和网络号。使用该表示法，让我们来看看路由表4，看起来将如何变化（表18-1）。

说明　表18-1与实际使用中的路由表相比已经相当简化了，如该表不包含子网掩码的域，还缺乏包含路由改急指示和表记录TTL值的域，这些域的用处以后会涉及。同时可以申明目的节点的完全规范网络地址以取代目的网络号。除此之外，正如已经提到的那样，这张表提供了传统格式的网络地址。这意味着本列中，网络地址不与任何特定网络协议相对应，无论如何，这张表包含了实际使用中的路由表的主要字段。

该表的第一个字段包含**分组目的地址**（packet destination address）。

在表的每一行中，目的地址后面是**下一个路由器的网络地址**（network address of the next router）。更精确地，这是为了将分组传送到目的地址必须途经的下一个路由器相应接口的网络地址。

表18-1 路由器4的路由表

目的地址	下一个路由器的网络地址	输出端口的网络地址	到目的网络的距离（跳数）
N1	IP_{12}（R1）	IP_{41}	1
N2	—	IP_{41}	0（直接相连）
N3	IP_{12}（R1）	IP_{41}	1
N4	IP_{21}（R2）	IP_{41}	1
N5	—	IP_{42}	0（直接相连）
N6	IP_{21}（R2）	IP_{21}	2
IP_B	IP_{21}（R2）	IP_{41}	2
默认	IP_{51}（R5）	IP_{42}	—

在将分组传送到下一个路由器之前，当前路由器必须获知分组要传送到哪个端口（IP_{41}还是IP_{42}）。为了这个目的，可以使用路由表的第三个域，其中包含**输出接口的网络地址**（network addresses of the output interface）。

网络协议的某些实现允许路由表中相同目的地址对应于多行，这种情况下，选择路由时会考虑**到目标网络距离的字段**（distance from the destination network field），根据网络分组中指定的规则（此规则常被称为**服务等级**（service class））这个距离被解释为任何度量。距离可以多种方式来衡量跳数、分组在通信链路上传输所需的时间或者选定路由上的特定链路可依靠特性。它还可以是与特定规则相关的反应路由质量的任何其他值。在表18-1中，网络之间的距离以跳数来衡量。对于直接与路由器端口相连的网络来说，可以认为这个距离是0，尽管有些实现从1开始计算该距离。

当在分组到达路由器时，IP实体从传送帧的头部检索出目的网络号并且将它与路由表中的每一行按顺序对比。匹配成功的行指定该分组转发时最近的路由器。举个例子，如果想要传送到网络N6的分组到达了路由器4的某一端口，路由表显示下一个路由器的地址为IP_{21}，这意味着转发过程的下一步，该分组会被传送到路由器2的端口1。

经常遇到的情况是，路由表指定了目的网络号而不是整个IP地址。因此，对于所有传递到同一网络的分组，IP会为它们安排同样的路由（当前我们并不考虑网络状态的变化，如路由器失效或者线缆中断）。然而，有些情况下，有必要为某网络节点选择**特定路由**（specific route），此特定路由与为所有其他网络节点指定的路由有所区别。为了做到这一点，有必要将此主机做为单独的一行添加到路由表中，此行必须包含主机的完整IP地址以及适当的路由信息。表18-1中有主机*B*的一条此类记录。比如说，假定路由器4的管理员，考虑到安全隐患，决定所有目的地为主机*B*（完整IP地址为IP_B）的分组都必须经过路由器2（接口IP_{21}）而不是路由器1（接口IP_{12}），因为经过路由器1的分组均被送往网络N3的其他主机。如果表中含有传送到整个网络以及送往某一主机的路由记录，当有送往该主机的分组到达该路由器时，将优先选择特定路由。

既然分组可能被送往互联网上的任意子网，那么好像每个路由表都必须保存互联网中的所有子网的记录。但是，大型网络中由于路由表会急剧膨胀而导致该解决方法效率低下，大型路由表的搜索工作也相当费时而且需要更多的存储空间。因此，在实际操作中，通过使用**默认路由**（default route）的特殊记录来减少路由表中的记录数是明智的。如果考虑到互联网的拓扑，你会发现位于互联网周边设施上路由器的路由表只能记录与其直接连接或尽头路由附近的网络号。至于其他网络，在路由表中插入一条记录，申明到达这些网络所经过某个路由器的路由路径，就足够了。此路由器即被称为**默认路由器**（default router）。路由表某一行必须含有一条特定的、名为**默认**（default）的记录，在我们的示例中，路由器4只为传送到网络N1～N6指定了路径。对于其他所有送往N7至N18的分组，该路由器为它们指定了路由器5的IP_{51}端口。

18.3.2　端节点上的路由表

路由问题除了中转节点（路由器）关心以外，也是终端节点（计算机）所关心的。对这个问题的处理早在安装于端节点上的IP检测分组是应该被送往另外一个网络还是被定位到本网络中的某一主机时就已经开始了。如果目的网络号与本网络号相一致，那么此分组无需被路由；否则，则需要路由。

端节点与中转节点的路由表结构类似，请再看下图18-2中所示的网络。属于网络N3的终节点*B*的路由表看上去如表18-2所示，这里，IP_B是计算机*B*的网络接口。基于这张表，计算机*B*在本地网络N3中的两个路由器中选择一个来传送特定的分组。

表18-2　计算机*B*的路由表

目的网络号	下一个路由器的网络地址	输出端口的网络地址	到目的网络的距离
N1	IP_{13}（R1）	IP_B	1
N2	IP_{13}（R1）	IP_B	1
N3	—	IP_B	0
N4	IP_{31}（R3）	IP_B	1
N5	IP_{13}（R1）	IP_B	2
N6	IP_{31}（R3）	IP_B	2
默认	IP_{31}（R3）	IP_B	—

端节点比路由器将默认路由技术贯彻得更加彻底，虽然它们也有可供配置的路由表，这些表的容量通常比较小。这是因为终端节点位于网络外围设备上。有时候端节点不配备路由表，这样，它们只能使用默认路由器里的地址信息。这种情况发生在组成网络中的众多端节点只有一个路由器时。然而即使存在多个路由器，端节点也必须选择正确的那个，通常制定默认路由以提高计算机的性能。

表18-3显示了互联网中的另外一个端节点（主机*A*）的路由表。该表的容量很小，意味着所有从节点*A*发送的分组必须经过路由器17的端口1或者永远不脱离网络N12的范围限制。该路由器在此路由表中，即被定义为默认路由器。

表18-3　端节点*A*的路由表

目的网络号	下一个路由器的网络地址	输出端口的网络地址	到目的网络的距离
N12	—	IP_A	0
默认	$IP_{17,1}$(R17)	IP_A	—

路由器与端节点之间行为的另外一个区别在于构建路由表的方法。按照规则，路由器通过交换控制信息自动创建路由表，与之相反的是，端节点的路由表总是由管理员手动创建。人工创建的路由表像普通文件一样保存在硬盘上。

18.3.3　搜索不含掩码的路由表

本小节描述路由器上IP所使用的路由表查找算法。当描述该算法时，我们使用表18-1和图18-2。

- 假定有分组抵达路由器的某一端口。IP从新到达的分组中检索出目的IP地址，为了说明方便，假设该分组指定了IP_B作为目的地址。
- 现在执行表查询的第一阶段（*first phase*），搜寻到目的主机的特定路由（*searching for a specific route to the destination host*）。将IP_B与路由表中的目的地址（*destination address*）按顺序一行一行进行对比，一旦匹配（如同表18-1中所示），就从输出接口（IP_{41}）标识符的同一行中找到下一个路由器的地址（IP_{21}），本步骤的查找过程完成了。
- 现在假定此表不包含目的地址为IP_B的行，也就是没有发现任何匹配。这样IP不得不进行表查

找的第二阶段（*second phase*），搜寻到目的节点的路由（*searching for the route to the destination address*）。从IP地址中检索出网络号（本例中，将从IP_6地址中检索出网络N3的网络号），然后重复表查询。此时，有必要找到与分组中指定网络号相匹配的网络号，如果发现匹配（与我们之前的例子相同），就从表中的合适行检索出下一个路由器的地址（IP_{12}）以及输出接口标识符（IP_{11}）。执行到这一操作，表查询就完成了。

- 最后，假定第一步和第二步均没有发现收到分组中目的地址的匹配，这种情况下，IP或者选择默认路由，即分组传递到IP_{51}地址，或者不存在默认路由时丢弃该分组。至此，表查询过程全部结束。

要点 这里很重要的一点，就是表查询的步骤被严格限定，但是路由表中各行的排列顺序，包括指定默认路由记录的位置，对结果并不产生影响。

18.3.4 不同格式路由表的例子

TCP/IP协议栈在实际应用时路由表的结构通常与前面考虑的路由表的简化结构相对应，然而特定IP路由表的格式依赖于TCP/IP栈的实现。现在考虑图18-3中路由器R1可操作的路由表的多个版本。

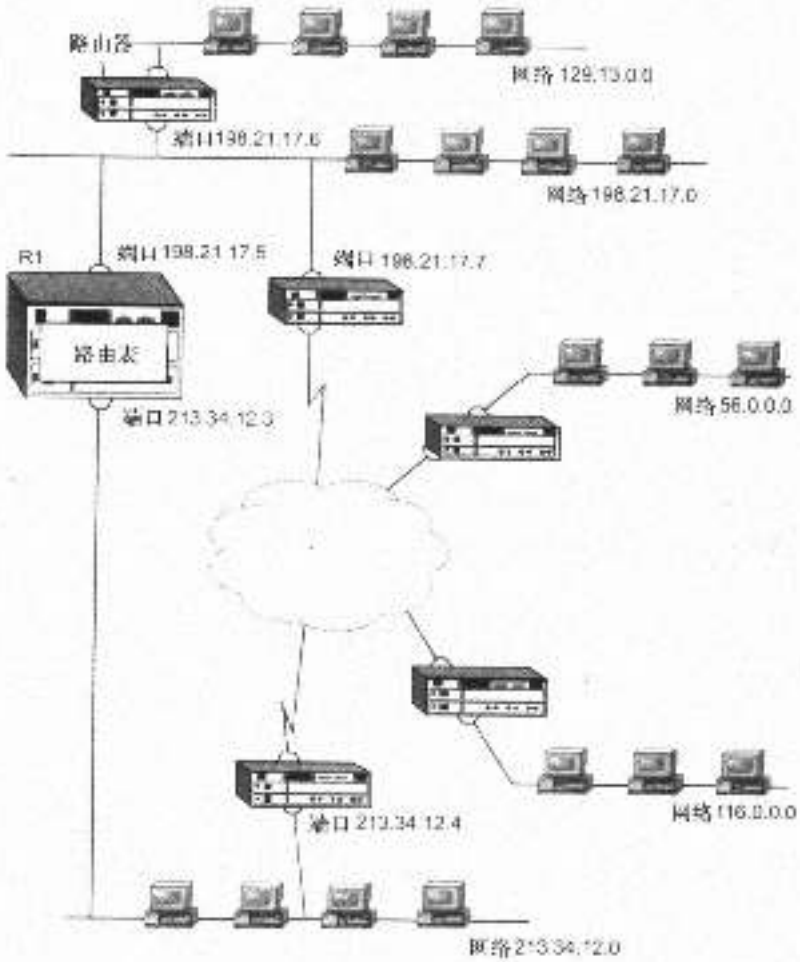


图18-3 被路由网络的示例

让我们从路由表的附带人工干涉并简化过的另外一种形式开始（表18-4）。这里存着到网络的三个路由（记录56.0.0.0、116.0.0.0以及129.13.0.0），直接相连网络的两条记录（129.13.0.0和213.34.12.0）以及默认路由的一条记录。

表18-4　路由器R1的简化路由表

目的网络地址	下一个路由器地址	输出接口的地址	到目的网络的距离
56.0.0.0	213.34.12.4	213.34.12.3	15
116.0.0.0	213.34.12.4	213.34.12.3	13
129.13.0.0	198.21.17.6	198.21.17.5	2
129.13.0.0	198.21.17.5	198.21.17.5	1
213.34.12.0	213.34.12.3	213.34.12.3	1
默认	198.21.17.7	198.21.17.5	—

工业制造行业的网络设备的路由表格式更加复杂。

如果你假定当前网络中的路由器R1是微软Windows 2000操作系统的内嵌软件路由器，那么它的路由表可能如表18-5中所示。

表18-5　Windows 2000内嵌路由器的路由表

网络地址	子网掩码	网关地址	接　口	度量值
127.0.0.0	255.0.0.0	127.0.0.1	127.0.0.1	1
0.0.0.0	0.0.0.0	198.21.17.7	198.21.17.5	1
56.0.0.0	255.0.0.0	213.34.12.4	213.34.12.3	15
116.0.0.0	255.0.0.0	213.34.12.4	213.34.12.3	13
129.13.0.0	255.255.0.0	198.21.17.6	198.21.17.5	2
198.21.17.0	255.255.255.0	198.21.17.5	198.21.17.5	1
198.21.17.5	255.255.255.255	127.0.0.1	127.0.0.1	1
198.21.17.255	255.255.255.255	198.21.17.5	198.21.17.5	1
213.34.12.0	255.255.255.0	213.34.12.3	213.34.12.3	1
213.34.12.3	255.255.255.255	127.0.0.1	127.0.0.1	1
213.34.12.255	255.255.255.255	213.34.12.3	213.34.12.3	1
224.0.0.0	224.0.0.0	198.21.17.6	198.21.17.6	1
224.0.0.0	224.0.0.0	213.34.12.3	213.34.12.3	1
255.255.255.255	255.255.255.255	198.21.17.6	198.21.17.6	1

如果你用某种常用硬件路由器取代路由器R1，那么同一网络的路由表可能看上去会不一样——如表18-6中所示。

表18-6　硬件路由器的路由表

目的地	掩　码	网　关	度量值	状　态	TTL	源
198.21.17.0	255.255.255.0	198.21.17.5	0	可用	—	连接
213.34.12.0	255.255.255.0	213.34.12.3	0	可用	—	连接
56.0.0.0	255.0.0.0	213.34.12.4	14	可用	—	静态
116.0.0.0	255.0.0.0	213.34.12.4	12	可用	—	静态
129.13.0.0	255.255.0.0	198.21.17.6	1	可用	160	RIP

最后，表18-7代表着同一个路由器R1在UNIX操作系统上实施软件路由器的路由表。

说明　既然网络结构和路由表之间没有明确的对应关系，那么可以为每个路由表的不同变化形式设计特定的版本，这些版本在特定网络的选定路由上可能会有所不同。这种情

况下，关注的重点转移到利用路由器的不同实施来表示路由信息形式的差异上。

表18-7 UNIX路由器的路由表

目的地	网关	标志	引用计数	使用	接口
127.0.0.0	127.0.0.1	UH	1	154	lo0
默认	198.21.17.7	UG	5	43 270	le0
198.21.17.0	198.21.17.5	U	35	246 876	le0
213.34.12.0	213.34.12.3	U	44	132 435	le1
129.13.0.0	198.21.17.6	UG	6	16 450	le0
56.0.0.0	213.34.12.4	UG	12	5 764	le1
116.0.0.0	213.34.12.4	UG	21	23 544	le1

虽然有着显著的差别，三种实际应用中路由表都含有分组路由IP所需要的所有关键数据。

这种数据列表包括**目的网络地址**（address of the destination network）（硬件路由和UNIX路由的目的字段、Windows 2000路由器的网络地址字段）。另一个路由表的必要字段是下一个路由器地址（硬件路由和UNIX路由的网关（*gateway*）字段、Windows 2000路由器的网关地址（*gateway address*）字段）。

第三个参数是**分组应该送往的端口地址**（address of the port to which the packet should be forwarded），有些路由表中直接指定该参数（Windows 2000路由器中接口（*interface*）字段），有些表中非显性地指定。比如说，UNIX路由器的路由表指定端口的传统命名而不是它的地址——le0代替端口地址198.21.17.5，le1代替地址213.34.12.3，lo0代替内部端口地址127.0.0.1。

在硬件路由器里，不存在以任意形式指定输出端口的字段。这是因为输出端口地址总是从下一个路由器的地址间接得到。比如，在表18-6的基础上尝试判断网络56.0.0.0输出端口的地址，从表中得出，此网络的下一个路由器地址为213.34.12.4，下一个路由器的地址必须属于直接与路由器相连的网络，此时为213.34.12.0网络。该路由器有连入那个网络的端口，端口地址213.34.12.3，可以在路由表第二行的网关（*gateway*）字段找到，此行描述了直接相连的213.34.12.0网络。对于直接连接的网络，下一个路由器的地址总是本地路由器的输出端口地址。因此，对于网络56.0.0.0，输出端口的地址将为213.34.12.3.

当前，在路由表的每一行中使用子网掩码字段是一种标准的解决方法。比如，考虑Windows 2000路由器的路由表（子网掩码（*netmask*）字段）以及硬件路由器的路由表（掩码（*mask*）字段）。在决定分组路由时路由器上的掩码处理在本章的后续会涉及。缺乏掩码字段意味着或者路由器只为三个标准地址分类服务或者为所有记录使用同样的掩码，这样降低了路由的灵活性。

既然每个目的网络在UNIX路由器的路由表中只提到一次，路由并没有太多选择。因此这种情况下，尺度是一个可选参数。其他表含有此字段；然而，它只是用于直接连接网络的一个指示符。因此，硬件路由器的路程0或Windows 2000路由器的路程1只是告诉路由器该网络是直接连入端口的。路程的其他值对应于远程网络，而路程为直接连接网络的选值（1或0）则是任意的。这里的关键点在于远程网络的路程必须由选定的初始值开始累加。

直接连接网络的指示符（*The indicator of the directly connected network*）告诉路由器分组已经到达了它的目的网络，因此，路由器不会将它传送到下一个路由器。取而代之的是，它将此分组直接传送到目的主机。因此，IP发起一个目的主机而不是下一个路由器IP地址的ARP请求。

然而，有些情况下路由器必须为每个远程网络记录存储路程值。当路由表中的记录为多个路由协议合作操作结果时，就会发生这种情况，如RIP。在这些协议中，任何远程网络上新收到的信息与表中的信息相对比，如果新的路程值比当前值小，那么新的记录将取代现存的。UNIX路由器

表没有路程字段，意味着此路由器不使用RIP。

标志只存在于UNIX路由器的表中，它们描述记录的下列特征：

- U——申明路由是活跃和可用的。硬件路由器的路由表中状态字段有着类似含义。
- H——显示某一主机的特定路由。
- G——申明分组路由经过中转路由器（网关），如果此标志遗失，那么此网络为本地连接。
- D——申明路由从ICMP的重定向（*redirect*）消息中获得，此标志只能存在于端节点的路由表中。如果它被设置，意味着端节点在之前传递分组的尝试中选择了下一个路由器，非最优的那个。该路由器，使用ICMP，规定将来所有传送到本网络的分组必须经由另外一个路由器。

在UNIX路由器的路由表中，存在着另外两个含有引用值的字段。引用计数（*refcnt*）字段申明此路由在分组传递过程中被引用的次数，使用字段申明此路由发送的字节数。

硬件路由器的路由表也有两个引用字段，在这种情况下，生存时间（*time to live*）（*TTL*）字段与分组的生存时间无关。如同很多其他表一样，这个字段防止记录的内容过期失效。TTL字段的当前值申明记录的TTL——即记录保持有效的时间长度（以秒为单位）。源（*source*）字段申明出现在路由表中的记录源自何处，尽管这个字段并不出现在所有的路由表中，对实际所有的路由器来说有三个主要记录源。

18.3.5　在路由表中记录的来源和类型

对实际所有的路由器，有三个（*three*）主要记录源。

- 路由表中的记录源之一为*TCP/IP*栈的软件实施（*the software implementing the TCP/IP stack*）。在路由器初始化时，此软件自动向路由表中插入几条记录，从而创建了**最小路由表**（minimum routing table）。

这类记录的列表包括**直接连接网络**（directly connected network）以及**默认路由**（default route）的信息，通常是手动配置计算机或路由器接口时输入的。在前面所举示例中，有网络213.34.12.0和198.21.17.0的记录，UNIX路由器上的默认路由记录以及Windows 2000路由器的0.0.0.0记录。

- TCP/IP软件也将特定地址的信息自动插入路由表中，在前面提供的示例中，Windows 2000路由器的路由表包含了最完整的此类记录集合。表中的多条记录与用于TCP/IP栈本地自测用的回环地址（*loopback address*）（127.0.0.0）相关，目的地址为224.0.0.0的记录是处理多播地址所必需的。除此之外，表中可能还包含处理广播的地址（比如，记录8和11包含了在恰当子网中广播消息的地址，并且此表的最后一条记录含有限制广播的地址）。注意有些路由表可能不包含特定地址的记录。
- 路由表中记录的第二个源为手动输入（*manual input*）。网络管理员利用一些特殊的网络设施直接生成这类记录，比如UNIX和Windows 2000提供的route命令。在硬件路由器中，也有特殊的指令来手动创建路由表中的记录。手动创建的记录通常是静态的，这意味着它们不存在过期问题，这些记录可能是持久的（一直保存直到重启路由器）或者临时的（存储在路由表中直到设备断电）。通常，管理员手动创建默认路由的记录，某些主机的特定路由记录也可以用同样的方法来创建。
- 最后，路由协议（*routing protocol*），诸如RIP或OSPF是路由表中记录的第三个源。这种记录总是动态的，意味着它们具有有限的TTL。

Windows 2000和UNIX操作系统的软件路由器并不在路由表中显示记录的源。而硬件路由器，与之相反，使用源（*source*）字段来实现这个目标。在表18-6的示例中，前两条记录由TCP/IP栈软

件依据端口配置数据而生成，在连接（*connected*）属性中显示。接下来两条记录被指定为静态（*static*），意味着它们是由管理员手动输入的。最后一条记录是RIP操作的结果，因此，它的TTL字段值为160。

18.3.6 不带掩码的IP路由的例子

图18-4示范了IP网络中，分组在互联网上转发的过程。在此例中，假定网络的所有主机都拥有基于分类的地址。我们将关注重点放在IP与ARP及DNS的交互上。

假设用户的计算机名为cit.mgu.com，位于以太网1中，需要建立到FTP服务器的连接。用户知道服务器的符号名称——unix.mgu.com。因此，用户使用下列命令访问FTP服务器：

> ftp unix.mgu.com

此条命令的执行分为三个阶段：

- 传递客户端发出的DNS请求以判断目的主机的IP地址
- 传递服务器的DNS响应
- 将分组从FTP客户端传到FTP服务器

现在让我们从第一步开始。

1. 传递DNS请求

1）FTP客户端向本机器上运行的DNS协议客户端部分递交请求。此模块反过来将格式化过的请求转给DNS服务器。此请求常见格式为："请问与符号名称unix.mgu.com相对应的IP地址是什么？"它被UDP分组封装，然后放入IP分组（图18-5）。DNS服务器的IP地址——203.21.4.6——存于此分组头部的目的地址中，因为此地址属于预先配置参数，故计算机的客户端软件可以获知。

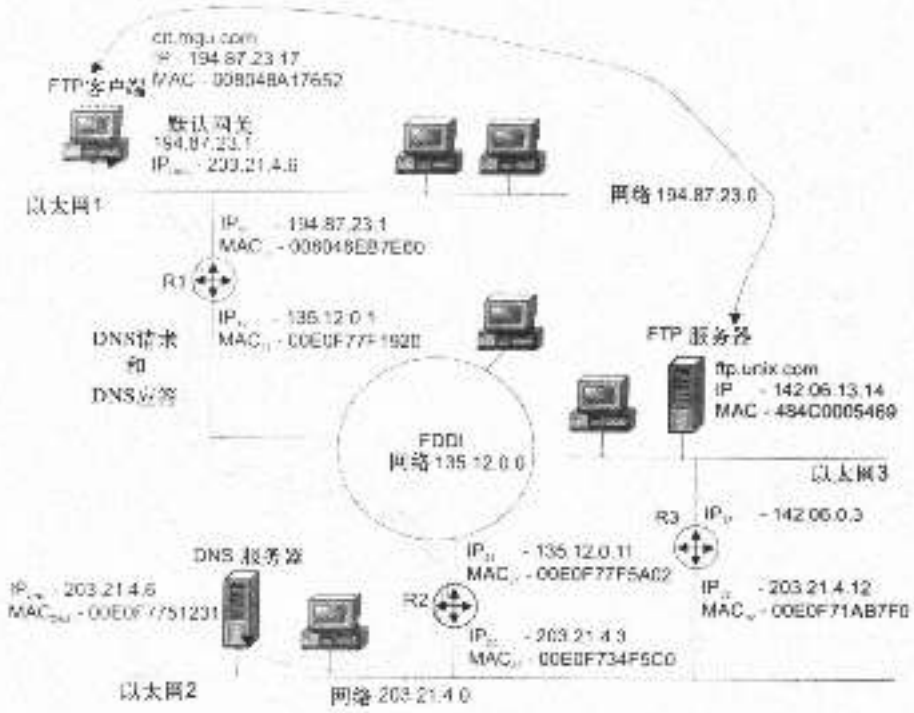


图18.4 IP路由示例

IP头部		UDP头部	DNS请求
IP发送端地址	IP接收端地址		
194.87.23.17	203.21.4.6		

图18-5　包含DNS请求的IP分组

2）在将IP分组封装成以太帧之前，有必要判断此分组应该在网间路由还是送往与发送端同一网络的某主机。为实现该目标，IP对比源地址和目的地址的网络号——比如，194.87.23.17和203.21.4.6。比较结果显示，分组应该送往另一网络；因此，必须被传送到最近的路由器。既然以太网1中只有一个路由器R1，那么此网络的所有终端节点均使用此路由器的地址——194.87.23.1——如同它是默认路由一样。这个地址也是客户端计算机预先配置的参数。

3）为了使以太网1能够将分组送往路由器R1，该分组必须被放置于以太帧的数据字段中并且提供一个MAC地址。这个问题由ARP通过搜索ARP表解决。如果表里不存在所需要的地址，客户端主机发送一条广播ARP请求："请问与IP地址194.87.23.1相对应的MAC地址是多少？"以太网1中的所有节点均收到此请求；然而，只有路由器R1的接口1做出响应："我的IP地址为194.87.23.1，并且我的MAC地址是008048EB7E60。"收到这条信息后，cit.mgu.com主机通过本地网络发出IP分组，此分组被封装进以太帧，其字段如图18-6所示。

以太网头部		IP头部		UDP头部	DNS请求
MAC发送端地址	MAC接收端地址	IP发送端地址	IP接收端地址		
MAC_0 008048A17652	MAC_1 008048EB7E60	194.87.23.17	203.21.4.6		unix.mgu.com?

图18-6　包含客户端计算机发送的DNS请求的以太网帧

FDDI头部		IP头部		UDP头部	DNS请求
MAC发送端地址	MAC接收端地址	IP发送端地址	IP接收端地址		
MAC_1 00E0F77F1920	MAC_2 00E0F77F5A02	194.87.23.17	203.21.4.6		unix.mgu.com?

图18-7　包含路由器R1发往路由器R2的DNS请求的FDDI帧

4）分组被路由器R1的接口1接收。以太网协议从帧中检索出IP分组并将其传送给IP。IP从分组中检索出目的地址：203.21.4.6，然后搜索本地路由表。假设路由器R1在路由表中有如下记录：

```
203.21.4.0          135.12.0.11   135.12.0.1
```

此记录表明发送往203.21.4.0网络的分组必须传递到路由器135.12.0.11，它位于连接到路由器R1接口135.12.0.1的网络中。路由器R1查询接口135.12.0.1的参数，发现一个与之相连的FDDI网络。因为FDDI网络中的MTU值比以太网的大，IP分组不需要分片。（MTU是存储于特定网络技术的传输单元数据字段中的数据报最大长度。）因此，路由器R1形成了FDDI格式的一个帧。

5）在这一步，路由器RI的IP实体必须通过已知IP地址——135.12.0.11判断出下一个路由器的MAC地址，它通过ARP实现此目标。假设，此时ARP表中有如下：

```
135.12.0.11          —          00E0F77F5A02
```

现在得知路由器R2的MAC地址（00E0F77F5A02），路由器R1将帧（图18-7）传入FDDI网络。

6）路由器R2上运行的IP实体以类似的方式工作。收到FDDI帧之后，它移去帧头，从IP头部检索出目的IP地址。然后搜索自己的路由表，从其中发现目的网络直接连接到它的第二个接口。

因此，它向以太网2发送下列ARP请求："请问与IP地址203.21.4.6相对应的MAC地址是什么？"收到响应DNS服务器的MAC地址——00E0F7751231之后，路由器R2将图18-8中的帧送往以太网R2。

以太网头部		IP头部		UDP头部	DNS请求
MAC 发送端地址	MAC 接收端地址	IP 发送端地址	IP 接收端地址		
MAC_{2r} 00E0F734F5C0	MAC_{DNS} 00E0F7751231	194.87.23.17	203.21.4.6		unix.mgu.com?

图18-8 包含路由器R2发送DNS请求的以太网帧

7）DNS服务器的网络适配器捕捉到以太网帧，发现其头部指定的目的MAC地址与它自己的MAC地址相匹配，于是把它送往自己的IP实体。在分析IP头部字段之后，IP从分组中检索出上层协议的数据。随后它递交DNS请求到DNS服务器的软件模块，DNS服务器搜索它的记录表，可能还会向其他DNS服务器提出请求。该操作的结果是，经DNS服务器格式化之后的响应，看起来大致如下："符号名称unix.mgu.com有对应的IP地址142.06.13.14"。

说明 分组在互联网中从客户端计算机传送到DNS服务器的整个过程中，IP头部字段里的源地址和目的地址一直保持不变。然而，每个携带分组在路由器之间传送的新帧，其头部字段的硬件地址在传送的每一步都有所改变。

2. 传递DNS响应

1）安装在DNS服务器上的TCP/IP栈将DNS响应封装进UDP数据报，然后封装进IP分组。注意，目的IP地址是从DNS请求中获知的。最后，IP得出结论，此分组需要的路由。

2）IP搜索路由表并且判断出下一个路由器的IP地址——IP22——203.21.4.3。

3）ARP判断出路由接口的MAC地址——00E0F734F5C0。

4）IP分组被放置在以太网帧的数据字段中，并且送往以太网2。

5）路由器R2收到帧并且执行步骤2和3描述的操作，随后它将FDDI帧送到路由器R1。

6）路由器R1从路由表中判断出收到的分组应该被送往与它的接口直接相连的网络。因此，IP请求ARP获取目的节点而不是路由器的MAC地址。

7）将目的地为FTP客户端的帧（图18-9）送往以太网1。

以太网头部		IP头部		UDP头部	DNS响应
MAC 发送端头部	MAC 接收端头部	IP 发送端头部	IP 接收端头部		
MAC_{1r} 008048EB7E60	MACo 008048A17652	203.21.4.6	194.87.23.17		*142.06.13.14*

图18-9 包含路由器R2发送的DNS响应的以太网帧

8）FTP客户端收到该帧并且从中检索出DNS响应。现在它可以继续执行指令，因为FTP服务器的符号名称已经被转换为IP地址形式。

> ftp 142.06.13.14

3. 从FTP客户端向FTP服务器传递分组

在网络上传输DNS请求和响应的步骤与前面描述的相似。然而，自己来描述该过程将是一个很好的练习。做这项练习时，请将注意力放在帧以及被封装分组中地址字段的值。

18.4　使用掩码的路由

随着越来越多的成员、掩码进入节点地址系统，路由算法变得越来越复杂。基于分类的寻址方法作为好的技术服务了多年后为何被遗弃？主要在于可分配的网络号短缺时，网络构建上的需求。

通常，当构建网络时，网络管理员会因为统一分配给他们的网络号数量不够而感到头疼。比如，在一个规划好的互联网里，不频繁交互的计算机分别被放置在不同的网络中。克服这种情况有两条途径，第一个是从某些集中权威机构那里获得更多的网络号，第二个是与使用掩码技术相关的更加普通的方法，它允许一个网络被分割成多个子网。

18.4.1　构造一个带同样长度掩码的网络

比如，假设一个管理员获得了B类地址：129.44.0.0，他可以组织一个从以下范围中选取主机号的大型网络：0.0.0.1——0.0.255.254，地址总数为$2^{16}-2$，因为由全0或全1组成的地址有特殊的含义，不适合主机寻址。可是，管理员并不需要单个无结构的网络。根据他们公司的需求存在着另外一种解决方法。根据此解决方法，公司网络必须被分成二个独立子网，每个子网的通信量在本地保持稳定，这种解决方案将简化网络诊断和维护，同时有助于为每个子网强制实行特定的安全策略。使用掩码分割大型网络的另外一个重要好处是，它允许公司网络对外部隐藏其内部结构，从而提高安全性。

图18-10显示了网络管理员将得到的整个地址空间划分为4个同等大小的块，每个包含2^{14}个地址。既然为主机取号的位数减少了2位，这四个网络的前缀都增加了2位，因此，这四个地址范围可以用带18个1或二进制形式的255.255.192.0掩码的IP地址来表示。



图18-10　将B类网络的地址空间分割成四个相等的块

```
129.44.0.0/18          (10000001 00101100 00000000 00000000)
129.44.64.0/18         (10000001 00101100 01000000 00000000)
129.44.128.0/18        (10000001 00101100 10000000 00000000)
```

129.44.192.0/18 (10000001 00101100 11000000 00000000)

从上面提供的记录中可看出，很明显管理员可以利用多余的两位为子网编号，这使得他可以从分配给他的一个大地址空间中创建出四个子网，本例中为子网129.44.0.0/18，129.44.64.0/18，129.44.128.0/18以及129.44.192.0/18。

说明 有些软件和硬件路由器，遵从先前的规范[RFC 950]，不支持由全1或全0组成的子网号，对于这种类型的设备，本例中使用的掩码255.255.192.0的网络号129.44.0.0是无效的，因为子网号的位值为00，基于同样的考虑，有着同样掩码的网络号129.44.192.0也将是无效的，此处，网络号为全1，然而，符合规范[RFC 1878]的现代路由器不受这些的束缚。

将一个大型网络拆分为四个同样大小的子网的示例见图18.11，所有从外网传入内部网络129.44.0.0的通信量都经过路由器R1传输。为了更进一步结构化信息流，另一个路由器R2被安装在内部网中，所有新创建的网络，129.44.0.0/18，129.44.64.0/18，129.44.128.0/18以及129.44.192.0/18，都连接在内部路由器R2的适当端口上。

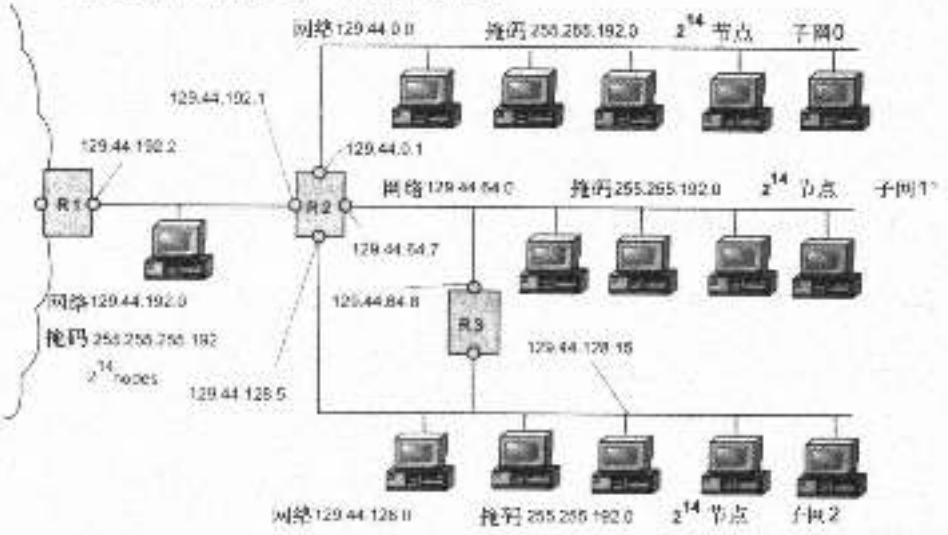


图18-11 使用相同长度掩码的路由

说明 其中一个网络，129.44.192.0/18，使用两个地址：129.44.192.1（路由器R2的端口）和129.44.192.2（路由器R1的端口），专取于创建外部与内部路由器的连接。另外两个地址：129.44.192.0和129.44.192.155具有特殊用途，此网络中的很多地址（$2^{14}-4$）都没有使用，当然，本例只是为说明将网络分成同等大小的子网是低效的而特别选取的。

从外部看来，此网络仍然是单个B类网络。然而，所有到达此网络的通信量都由本地路由器R2分配到四个子网中。当分类机制不起作用时，路由器必须有其他机制来判断目的地址字段的32位数中哪个部分代表着网络号。路由表（表18-8）包含了一个额外的掩码字段来实现此目的。

表18-8 带有相同长度掩码的网络中路由器R2的路由表

目的地址	掩 码	下一个路由器的地址	端口地址	状 态
129.44.0.0	255.255.192.0	129.44.0.1	129.44.192.2	连接
129.44.64.0	255.255.192.0	129.44.64.7	129.44.64.7	连接
129.44.128.0	255.255.192.0	129.44.128.5	129.44.128.5	连接

（续）

目的地址	掩　码	下一个路由器的地址	端口地址	距　离
129.44.192.0	255.255.192.0	129.44.192.1	129.44.192.1	连接
0.0.0.0	0.0.0.0	129.44.192.2	129.44.192.1	—
129.44.128.15	255.255.255.255	129.44.64.8	129.44.64.7	—

表中的前四条记录对应着直接连到路由器R2端口的内部子网。

带掩码0.0.0.0的记录0.0.0.0对应的是默认路由。

其后一条记录申明了到主机129.44.128.15的特殊路由。对于申明了主机完全IP地址的表记录，其掩码为255.255.255.255。与129.44.128.0网络的所有其他节点相比，来自路由器R2接口129.44.128.5送往该主机的分组将经路由器R3传输进来。

18.4.2　考虑掩码的表查找算法

搜索带掩码路由表的算法与前面描述的不带掩码的表算法有很多相同之处。然而，它也有着重要的改变：

1）通过检索分组的目的地址（*retrieving the packet's destination address*），IP开始为新到达的分组搜索下一个路由器。为了以示区别，将它指定为IP_D。然后，IP启动路由表搜索的算法，其步骤与不带掩码的表搜索步骤相似，也包含两个阶段。

2）第一阶段包括向特定路由搜索IP_D地址（*The first phase consists of searching for a specific route for the IP_D address*）。为了实现此目标，IP从每个掩码值设为255.255.255.255的表记录中检索目的地址，检索到的地址与IP_D分组中的目标地址相对比。一旦发现匹配，即从此记录中取出下一个路由器的地址。

3）当搜索了整个表仍然没找到匹配的地址时，就执行第二阶段（*second phase*）。在这一阶段中为带IP_D地址的分组所关联的主机群搜索一个通用路由，为做到这一点，IP再次搜索路由表。

4）对每个下一条记录（*next record*）执行下列动作：

- 对从分组中检索出的目的地址“应用”当前记录中的掩码（M）到IP_D与M。
- 将结果与存储在路由表中同一记录的目的地址字段里的值进行比较。
- 如果找到匹配，IP以适当的方式标记出此记录（*marks this record in an appropriate way*）。
- 若尚未遍历过所有的记录，IP将处理下一条（返回到步骤4）。若所有的记录都已经处理过，包括含有默认路由信息的那条，则协议继续处理第5步。

5）遍历过整个路由表之后，路由器执行下列动作中的一个：

- 若没找到任何匹配，并且也没有默认路由，则此分组被丢弃。
- 若找到一个匹配，此分组根据含有匹配地址的记录指定的路由进一步转发。
- 若找到多个匹配，则协议比较所有标记的记录并且选取一个匹配位数最多的记录所指定的路由。换句话说，当分组指定的目的地址属于多个子网时，路由器选用最明确的路由（*the router uses the most specific route*）。

说明　在很多路由表中，地址0.0.0.0和掩码0.0.0.0的记录对应着默认路由。传入分组中的任何地址，在应用了掩码0.0.0.0之后，会形成网络地址0.0.0.0，其与记录中指定的地址相对应。既然掩码0.0.0.0长度为0，此路由被认定为最不明确的并且只在路由表中无其他匹配时使用。

让我们来看一下路由器R2（图18-11）如何利用前面描述过的算法来操作它的路由表（表18-8）。假定目的地址为129.44.7.200的分组抵达路由器R2，则该路由器上的IP实体会首先将该地址与地

址129.44.128.15相对比，后者已经定义了一条确切的路由。既然没找到任何匹配，IP会顺序查找表中的所有行，应用掩码并对比结果直至找到目标地址和表中行的匹配网络号。至此，该分组的路由已经找到：它必须被传送到输出路由端口129.44.64.7以及直接连接到该路由器的129.44.64.0网络。

18.4.3 使用可变长的掩码

很多情况下，将网络分割成容量不同的子网效果更好。特别是，根据“点对点”方式连接到两个路由器的子网，即使C类网络时，其可用的地址数也太多了。

图18-12显示了与前面所述例子中相同地址空间的分配示例，也就是，129.44.0.0/16。一半可用地址（2^{15}）被用来创建网络号129.44.0.0和掩码为255.255.128.0的网络。下一段地址，整个地址空间的1/4（2^{14}），被分配给掩码为255.255.192.0的129.44.128.0网络。

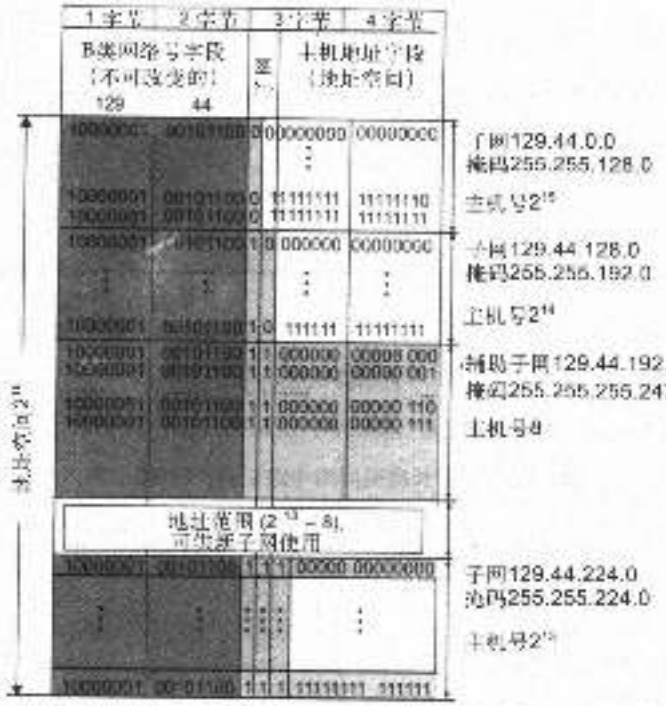


图18-12 利用变长掩码将B类网络（129.44.0.0）的地址空间分割成容量不同的子网

之后，从地址空间中“切除”一小段，旨在创建一个连接内部路由器R2到外部路由器R1的网络。为这样的一个退化网络中的主机编号，分配2位足够了，主机号的四个可能组合——00, 01, 10和11——00和11有着特殊意义不能分配给主机。其余两组，10和01，已经足够路由器端口寻址使用。这种情况下主机号码字段为2位长，二进制表示的掩码为：255.255.255.252，如图18-12中所示，网络号为129.44.192.0。

> 说明 由于除了两个路由器端口以外没有主机可以连入这种网络，所以不需要为路由器之间点对点的连接分配IP地址。大多数情况下，管理员会选择给这种网络分配一个IP地址，比如，为了隐藏网络的内部结构而这么做。这通过利用网络地址转换（NAT）技术，根据输入路由器端口地址来访问此网络来实现。除此之外，利用IP网络辅设通信量隧道可能也需要此地址，比如说，可以用这个方法来创建IPX或者加密IPSec通信量隧道，隧道和NAT将在第20章中描述。

余下的地址空间根据公司的需求可以被“分割”成任意大小的任意多个网络。比如，管理员可以从余下地址池（$2^{14}-4$）中创建一个足够大的包含2^{13}个主机的网络，余下的主机号几乎同样多（$2^{13}-4$）。另一方面，可用这些地址来创建新的网络。比如，这个“余者”可以被用来创建31个网络，其中每个在规模上都等同于一个C类网络，还可以用来创建多个更小的网络。很显然，应该选择另外一个分法。然而，当使用可变长掩码时，显然管理员能够拥有更加有效地使用所有可选的地址。

图18-13显示了使用可变长掩码构建的网络。

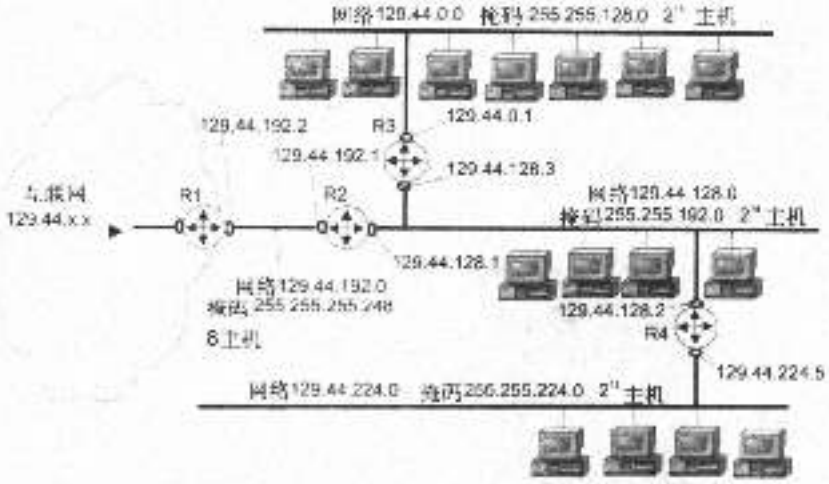


图18-13　利用变长掩码构建网络

考虑下路由器R2如何处理传送到它接口的分组（表18-9）。

表18-9　可变长掩码网络中路由器R2的路由表

目的地址	掩　码	下一个路由器的地址	输出地址	距　离
129.44.0.0	255.255.128.0	129.44.128.3	129.44.128.1	1
129.44.128.0	255.255.192.0	129.44.128.1	129.44.128.1	连接
129.44.192.0	255.255.255.248	129.44.192.1	129.44.192.1	连接
129.44.224.0	255.255.224.0	129.44.128.2	129.44.128.1	1
0.0.0.0	0.0.0.0	129.44.192.2	129.44.192.2	—

假设到达路由器R2的分组的目的地址为129.44.192.5。因为此表不包含明确路由，所以路由器进行第二阶段的表搜索，也就是，顺序分析所有的行以找到一个匹配的目的地址：

（129.44.192.5）AND（255.255.128.0）= 129.44.128.0——没有匹配的目的地址

（129.44.192.5）AND（255.255.192.0）= 129.44.192.0——没有匹配的目的地址

（129.44.192.5）AND（255.255.255.248）= 129.44.192.0——匹配目的地址

（129.44.192.5）AND（255.255.224.0）= 129.44.192.0　没有匹配的目的地址

因此，有一行中存在着匹配。分组将被送到与当前路由器直接连接的网络——送往129.44.192.1输出接口。

如果地址为129.44.192.1的分组从外部网络抵达，同时路由器R1未使用掩码，那么分组将被送往路由器R2且同时返回到中转网络。这样做显然效率低下。

如果路由器R1表中的所有路由都使用可变长掩码，那么路由将更加有效。这种表的分段见表18-10所示，表中两条记录中的首条指定目的地址以192.44打头的所有分组都必须被送往R2路由

器R2。此记录为基于129.44.0.0的所有子网执行**地址聚合**（address aggregation）。第二条记录指定在129.44.0.0的所有可能的子网中，有一个网络，129.44.192.0/30，分组可以直接传送给它而无需通过路由器R2转发。

说明 在使用掩码机制时，只有目的IP地址在IP分组中传递——没有目的网络的掩码。因此，不可能判断出所接收分组的IP地址，哪部分是网络号、哪部分是主机号。如果所有子网掩码都是同样大小，这不会有任何问题。然而，如果子网使用的是可变长掩码，那么路由器必须有判断哪些掩码与哪些网络号相对应的机制。可以使用路由协议来达到此目标。路由协议携带网络地址的信息以及与路由器中这些号码相对应的网络掩码信息。这种协议包括RIPv2和OSPF。至于RIP,它不携带网络掩码。因此，该协议不适合使用可变长掩码。

表18-10 对路由器R1的路由表进行分段

目的地址	掩 码	下一个路由器的地址	端口地址	距 离
....	....	....	....	....
129.44.0.0	255.255.0.0	129.44.192.1	129.44.192.2	2
129.44.192.0	255.255.255.252	129.44.192.2	129.44.192.2	连接
....	....	....	....	....

18.4.4 复用地址空间

管理员并不是在开始配置网络接口和创建路由表时才意识到掩码管理的复杂性，早在网络规划阶段就已经要面临这个问题了。网络规划包括决定公司范围内各子网的网络号、衡量它们所需的地址数、从提供商获取地址池、以及将可用的地址空间分配给这些子网。事实证明网络规划并不是一个无关紧要的任务，特别是在地址短缺的条件下。

现在让我们来看一下如何利用掩码来组织**覆盖地址空间**（overlapping address space）。

假设某公司的管理人员决定申请足够创建一个网络的地址池，其结构显示于图18-14中。客户端网络包括三个子网，其中的两个是内部部门级别的子网：以太网容纳600用户，令牌环网络容纳200用户。公司同时提供了包含十个主机的独立网络，旨在为公共访问模式的潜在客户提供信息。对于公司范围的网络，其中有网络服务器、FTP服务器和其他公共信息的提供商，这种分区通常叫做**非武装区**（DMZs）（demilitarized zone）。而且，还需要一个连接到服务提供商的只含有两个主机的网络，因此，分配网络接口所需要的地址总数为812。除此之外，有必要确保可用地址池中包含由全1和全0组成的广播地址。因为任何网络中所有主机地址都有着同样的前缀，很明显为了组建这样的网络所需客户端的最小地址数可能与812相差甚远，后者仅仅是简单相加而得。

本例中，提供商决定分配给客户端一段包含1 024个地址的连续地址池。选择1 024是因为它是与所需地址数最贴近的2的指数（2^{10} = 1 024）。提供商在地址空间中找到此长度的一段区域分配给它——131.57.0.0/16，注意此空间的一部分已经分配给其他客户端，正如图18-15所示。分配给客户的地址区域分别用S1、S2、S3来表示，提供商在尚未分配出去的可用空间中找到一段连续空间，容量为1 024个地址，区域的起始地址是其大小的倍数。因此，新的客户端将获得地址池131.57.8.0/22，图示中用S来表示。

随后，到了网络规划中最难的一步，需要将从提供商获得的地址池S分配给公司的四个子网。首先，管理员决定将完整的地址池131.57.8.0/22分配给最大的子网，它包含600个节点的以太网（图18-16）。分配给此子网的网络号与从提供商获得的网络号相一致，然后，管理员该如何处理余

下的二个子网呢？管理员记得以太网仅仅需要600个地址，在余下的424个地址中，管理员"抓出"256个分配给令牌环网。鉴于令牌环网只需要200个地址，管理员又将其"切"为两块：用来组织DMZ的16个地址131.57.9.16/28网以及连接公司与ISP网络的4个地址组成的131.57.9.32/30网。最后，该公司的所有子网都被获得了足够多（有的甚至是冗余）的地址数。

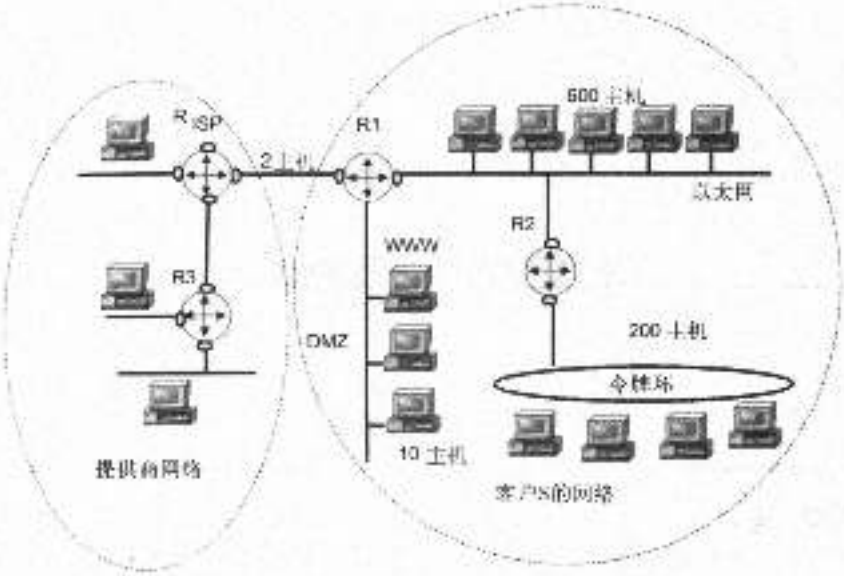


图18-14　提供商和客户端网络

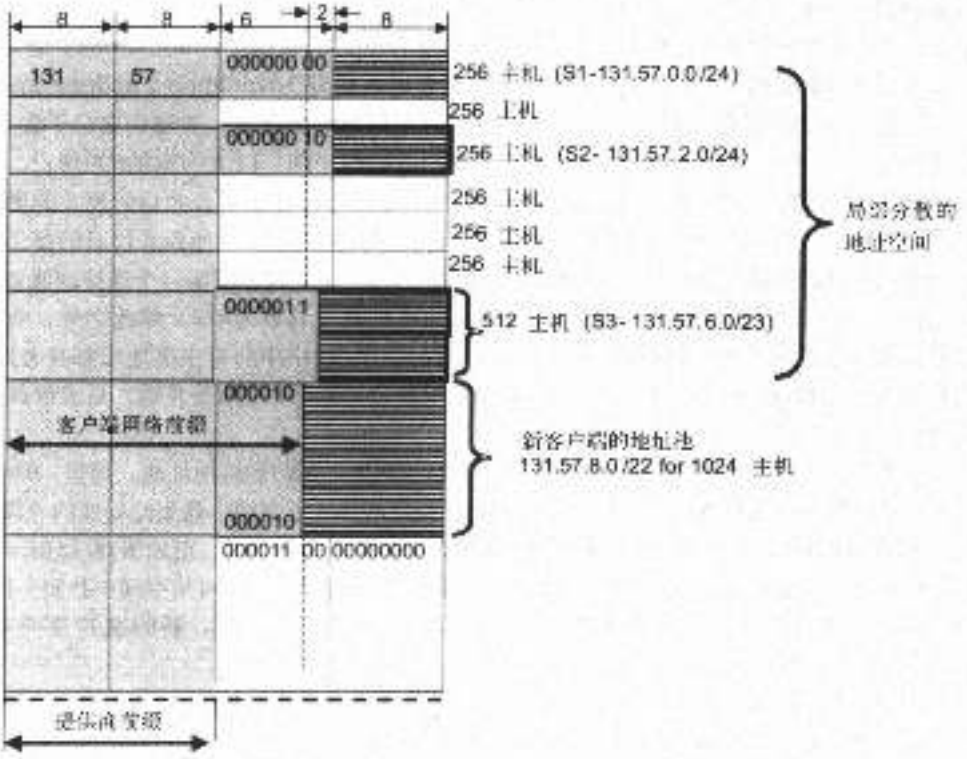


图18-15　提供商的地址空间

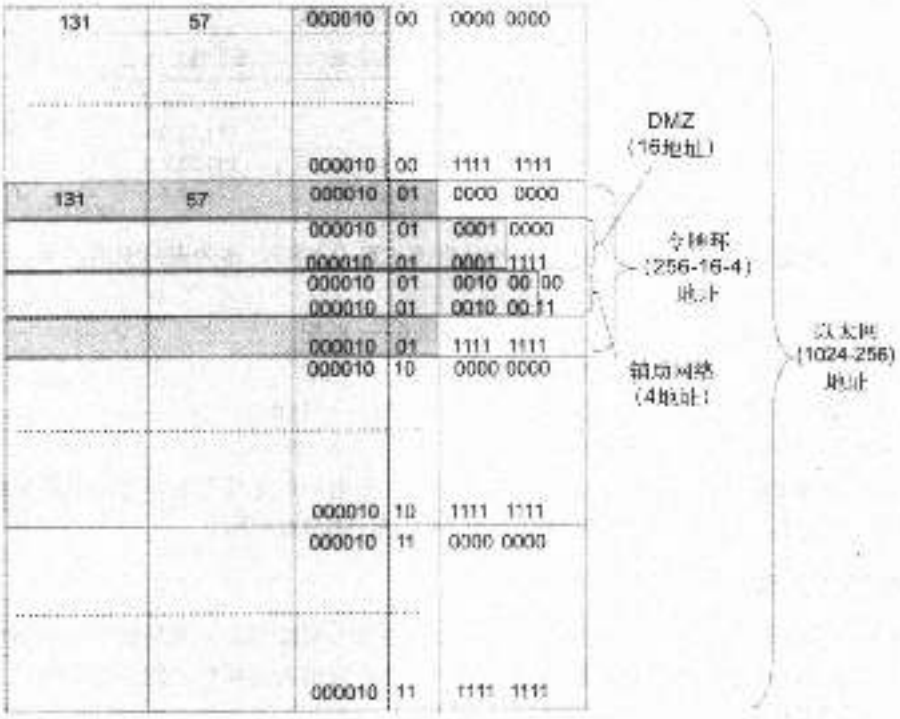


图18-16 为客户的网络规划地址空间

下一阶段该配置终端节点和路由器的网络接口了，为每个接口配置IP地址和相应的子网掩码。图18-17显示了为客户配置好的网络。

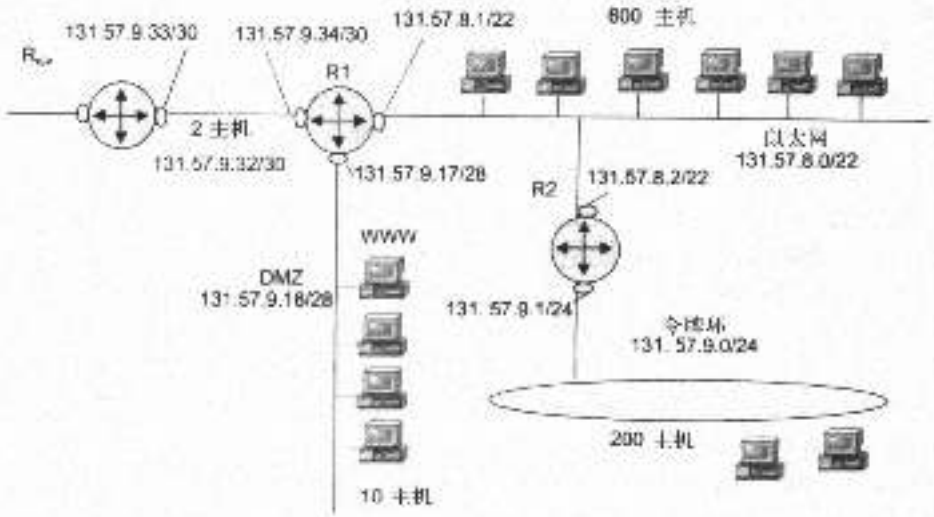


图18-17 配置客户的网络

配好了所有的网络接口之后，需要为客户的路由器R1、R2创建一个路由表，它们可以自动生成或者由管理员手动创建。表18-11显示了路由器R2的路由表。

此表中没有默认路由，这意味着发往本网的所有分组，如果其地址在表中不存在的话，将无一例外地被路由器丢弃。

表18-11　路由器R2的路由表

目的地址	掩　码	下一个路由器的地址	输出接口地址	距　离
131.57.8.0	255.255.252.0	131.57.8.2	131.57.8.2	连接
131.57.9.0	255.255.255.0	131.57.9.1	131.57.9.1	连接
131.57.9.16	255.255.255.240	131.57.8.1	131.57.8.2	1
131.57.9.32	255.255.255.252	131.57.8.1	131.57.8.2	1

举个例子，假定目的地址为131.57.9.29的分组抵达路由器R2，作为表搜索的结果，可以获得下列结果：

(131.57.9.29) AND (255.255.252.0) = 131.57.8.0——匹配

(131.57.9.29) AND (255.255.255.0) = 131.57.9.0——匹配

(131.57.9.29) AND (255.255.255.240) = 131.57.9.16——匹配

(131.57.9.29) AND (255.255.255.252) = 131.57.9.28——不匹配

根据路由表的搜索算法，若存在多个匹配，将选择分组目的地址与表中行的目的地址匹配长度最长的路由。因此，地址为131.57.9.29的分组将被传递到DMZ网络。

18.4.5　路由和CIDR

最近几年，因特网中发生了很多变化：主机和网络数大幅度增长、通信强度大大增加并且传输的数据类型也有所改变。由于路由协议的不完善，携带路由表更新信息的消息交换在主干路由器上开始出现失败的情况，这是因为处理大量控制信息引起的拥塞所至。比如，主干因特网路由器的路由表会包含上百甚至上千个路由。

为了克服该问题，**无类别域间路由（CIDR）（classless interdomain routing）技术**应运而生。

CIDR的主要思想如下：必须为每个ISP分配IP地址空间里连续的一段，使用这种解决方法时，每个服务提供商的所有网络地址将有着相同的高位部分—前缀（prefix）。因此，主干因特网中的路由可以在前缀的基础上实现，而不是完全规范的网络地址。这意味着每个网络只有一条记录，而不是多条，每个拥有匹配前缀的网络只要一条记录就足够了。地址整合会减少所有层路由器的路由表容量，从而使得路由表处理得更快且扩大因特网带宽。

在本章前面提到过的示例中，公司范围网络的管理员使用掩码将从ISP得到的一段连续的地址池拆分为多个部分来构建网络。这种使用掩码的方法叫做**子网化（subnetting）**（RFC950）。

同时，使用掩码将网络拆成子网也有着相反的效果——**网络聚合（network aggregation）**。简单来讲，为了路由所有进入公司网络的流量，其中公司网络被拆成了多个子网，只需要在外部路由器上保存一条记录就足够了。此记录的目的地址必须指定为这些网络所共有的前缀。前缀的右界限使用恰当的掩码来指定。这就是**超网化（supernetting）**，一个拆子网化相反的操作。超网化意味着掩码被用于将几个子网组建成单个大型网络。

回到图18-15，图中显示了包含S1、S2、S3、S的ISP地址空间，这四个地址区域分别分配给四个客户端。该示例在图18-18中也有所体现。作为表18-12中聚合客户网络的结果，将为每个客户端分配一行，无论它们在网络中组织的子网数是多少。比如，为客户S的七个网络并不会提供四个路由，取而代之的是，只为这些网络提供一个路由。

对于支持使用$R_{external}$路由器客户端的顶层提供商，并不会留意到本地提供商拆分地址空间的举动。带有掩码255.255.0.0的记录131.57.0.0完全描述了$R_{external}$路由器中的本地提供商的网络。

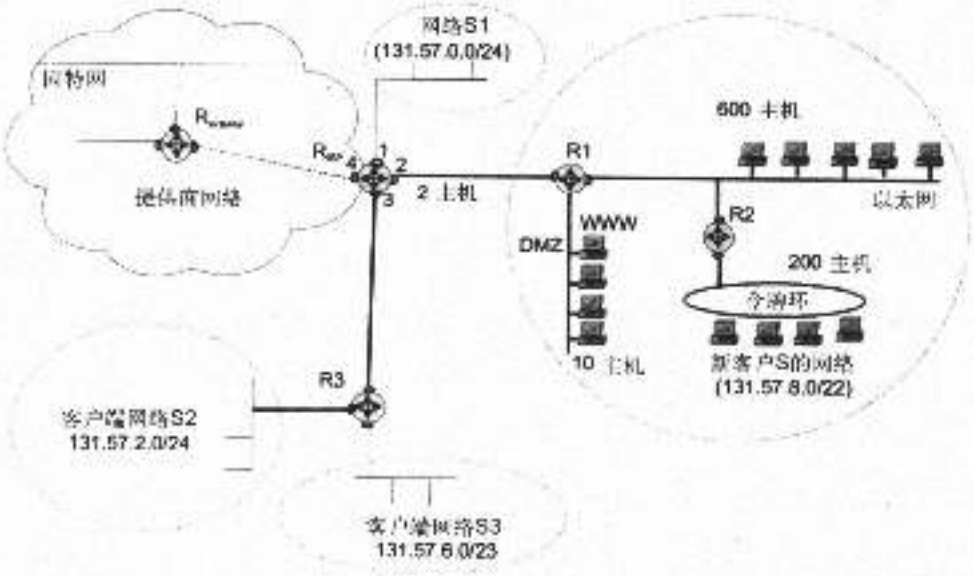


图18-18 超网化

表18-12 R_{BP}路由器的路由表

目的地址	掩 码	下一个路由器	输出接口号	距 离
131.57.0.0 (S1)	255.255.255.0	—	1	直接
131.57.2.0 (S2)	255.255.255.0	R3	3	1
131.57.6.0 (S3)	255.255.254.0	R3	3	1
131.57.8.0 (S)	255.255.252.0	—	2	1
默认	0.0.0.0	$R_{default}$	4	—

因此，对CIDR技术的介绍包括下列解决方法：

- 更加节约利用地址空间。利用CIDR技术，提供商可以根据客户需求“切割”分配给它们的地址空间。与此同时，客户端保留一些地址已备将来使用。
- 由于路由聚合而造成路由表中记录数的减少。因此，路由表中单个记录可以代表着大量子网，如果所有的ISP都遵从CIDR策略，在骨干路由器上的增益是非常显著的。

地址本地化是CIDR有效使用的必备条件。地址本地化意味着拥有匹配前缀的地址被分配给地理位置相邻的网络，只有实现了这一点，流量聚合才成为可能。

很不幸的是，地址分配和很多时候是任意的。重新分配网络号是解决此问题的一个基本方法，然而，这一步骤耗费大量的时间和财力。因此，必须以某种方式督促用户执行它，这种督促包括为路由表中每一行或者根据网络中的主机数收费。第一个需求使客户有意识地从供应商处获取允许流量根据网络前缀路由到他所处网络的地址。之后，携带公司网络号的记录将不会再出现在主干路由器上。要为每个主机地址付费的需求同时也促进用户重新计算并且不会获取超过所需的地址数。

CIDR技术成功运用于IP的当前版本，IPv4，同时也被路由协议OSPF，RIPv2，BGP4（主要用在因特网主干路由器上）所支持。在IP的新版本，IPv6，中使用CIDR的特殊特性将在本章节后面提到。

18.5 IP分组的分片

IP的一个重要特性，与所有其他网络协议（如IPX）的区别，在于当在网络间传递分组时它可

以执行**动态分组分片**（dynamic packet fragmentation）。在这些网络的帧数据字段最大长度（MTU）有差异时，就需要用到分片了。IP特有的分组分片能力在很多方面提升了TCP/IP技术的灵活性。

18.5.1　MTU作为一个技术参数

首先，请允许我指出在源节点中消息分片和网络的中转节点——路由器中的动态消息分片是不同的。在所有的协议栈中，都存在着负责分割应用层消息的协议，将消息分成适用于数据链路层帧的分片。为达到此目的，它们分析下层技术的类型并且为之定义MTU。

在TCP/IP栈中，该问题由TCP解决，它将从应用层传输过来的字节流分割成所需大小的段（比如说，当网络下层协议使用的是以太网协议时为1 460个字节）。因此，**发送端**（sender）的IP通常不会使用分片功能。然而，当需要从当前网络向另外一个MTU值小一点的网络传送分组时，对于路由器来说情况就不一样了。这种情况下，就要使用IP分片功能。

从表18-13中看出，很明显多数流行技术的MTU值有着很大的差异，这意味着在当前异构网络中分片是一个常用的操作。

表18-13　典型的MTU值

技　术	MTU
DIX以太网	1 500字节
以太网802.3	1 492字节
令牌环（IBM，16Mb/s）	17 914字节
令牌环（802.5, 4Mb/s）	4 464字节
FDDI	4 352字节
X.25	576字节

18.5.2　分片参数

分片的主要思想是将到达网络的MTU值偏大的分组拆成短小的分组，**分片**（fragment），当分组将转发给MTU值小的网络时。在分片在网络中传输时，它们会在某些中转路由上再次分片，必须为每个分片提供一个全值的IP头。

有些头部字段，比如标识符（*identifier*）、*TTL*、*DF*和*MF*标志（*flags*）以及偏移量（*offset*），都是直接服务于将分片重组成源消息的过程。

- 分组的接收端使用标识符字段来识别同一个分组的所有分片（*identifier field for recognizing all fragments of the same packet*）。发送分组的IP模块用对当前“发送端-接收端”对来说唯一的值来填充标识符字段。在此分组存在于IP互联网中的全程都必须符合该条件，为了确保符合条件，发送分组的IP实体可以跟踪指定的标识符。比如说，可以通过维护一张表来实现，表中每条记录与各个建立连接的目的主机相关。这种表的每条记录包含IP网络中分组的TTL最后取值。但是，因为标识符字段允许65 536个不同的值，有些IP实施随机选取此范围中的值，它在整个分组传送中保持唯一的可能性是非常大的。
- 发送端指定分组存活于网络中的时间TTL。
- 分组分片的偏移量（*offset*）字段通知接收端此分片在源分组中的位置。因此，第一个分片总是拥有0偏移值。若分组没有被拆分，其偏移值也为0。
- *MF*标志设为1表明着刚到达的分片不是最后一个，发送一个未被分片的分组时，IP实体将MF标志设为0。
- *DF*标志设为1表明在任何条件下当前分组都不允许分片，如果标为不能被分片的分组无法在不分片的情况下到达目的节点，那么IP实体将丢弃它同时向发送端发送一条恰当的ICMP消息。

说明　阻止分组分片在某些情况下有助于让应用程序运行得更快，为了实现这个目的，首先有必要调查网络并断定可沿着整个路途传送而无需分片的分组的最大长度。然后，只有不超过该大小的分组才被使用。该特性也可以被用于在目的主机的资源不足以组装分片是防止分片。

18.5.3 分片和组装分组的过程

首先考虑**分片**（fragmentation）的过程。在将新抵达的分组分割成分片之前，安装在路由器上的IP为新的分片分配多个缓存。

然后从源分组的IP头部中拷贝多个字段内容存入这些缓存，从而创建出新分片分组的IP“哑”头部。IP头部的某些参数被拷进所有分片的头部；其他只保留在第一个分片的头部。分片过程会改变分段IP头部的某些字段值。因此，每个分片都有它自己的头部校验和的值、分片偏移地址，以及分组总长度。除了最后一个分片外，其他所有分片的MF标志都被设为1；在最后一个分片中，它被设为0。

每个分片的数据字段内容是通过分割源分组的数据字段内容形成的，进行此操作必须满足下列两个条件：首先，分片大小（IP头部加上数据字段）不能超过其下层技术协议的MTU值；其次，除了最后一个分片以外，每个分片的数据字段大小必须是8字节的倍数。最后一部分的数据大小等同于剩余数据的大小。

现在来考虑一下被分片的分组如何重新组装，此过程发生在目的主机上。

说明 注意到IP路由器并不会将分组分片组装成更大的分组，即使在它们的传送路径中有些网络允许这种聚合。这样做的理由很明显：同一消息的不同分片可能会采用互联网中不同路由。因此，并不能确保所有分片都可以通过同一路由器。

因此，对于每个被分片的分组来说，目的主机为它分配一块特殊的缓存，接收IP实体用以存放携带匹配源地址、目的地址以及标识符和协议字段值的IP分片。所有这些属性指定这些分组都是同一源分组的分片，组装过程包括将每个分片的数据放入分组头部偏移地址字段指定的位置上。

当源分组的第一个分片抵达目的主机时，该主机启动组装计时器以计算允许等待其他分片抵达的最晚时间。IP的不同实现应用不同规则来进行超时期限的选择。比如，计时器可以根据RFC的推荐被设为固定值（60秒到120秒），原则上这段时间足够将分组从发送端传到接收端。而其他实现可以使用适当的算法来决定这个期限，该算法衡量网络中的时间并统计地估算分片到达的预期时间。最后，在所接收到分片携带的TTL值的基础上选择超时。后面一个解决方案建立在当接收到的分组TTL过期时，等待其他分组就毫无意义的思想上。

说明 如果分组至少有一个分片尚未准时抵达目的主机（准时意味着在计时器过期之前），那么将不会为丢失的分片采取任何行动，直接丢弃所有已收到的分片。目的主机发送一个ICMP错误信息给发送端。IP的这种行为对应于它的“尽力服务”计划（即，协议尽自己的最大可能发送分组但是并不提供担保）。

数据字段中空白间隙（“漏洞”）和最后分片的抵达（MF=0）意味着组装过程完成。数据被组装之后，才有可能传往上层协议，比如TCP。

18.5.4 分片的例子

考虑一个在路由器上进行分片的示例（图18-19）。

假定发送端主机连接在MTU为17 914字节的网络上，可以把它当成一个令牌环网络。原则上传输层知道下层技术的MTU并且以恰当的方式选择分片大小。假设此例中，6 600字节的消息从传输层传递到IP层，IP在此消息的基础上形成IP分组的数据字段，并为之提供了头部。请格外留心与分组分片相关的头部字段是如何填充的。首先，分组被指派一个唯一的标识符，如，12456。其次，既然分组还没有被分片，那么偏移量（offset）字段设为0，*MF*标志也被设为0，表示后续不存在分片。第三，*DF*标志被设为1，意味着该分组可以分片。IP分组的总长度为6 600 + 20（IP头

部大小）——也就是6 620个字节，该分组的总长度符合令牌环帧的数据字段。然后发送端主机的IP实体将这个帧传送到它的网络接口，由接口负责将帧送往下一个路由器。

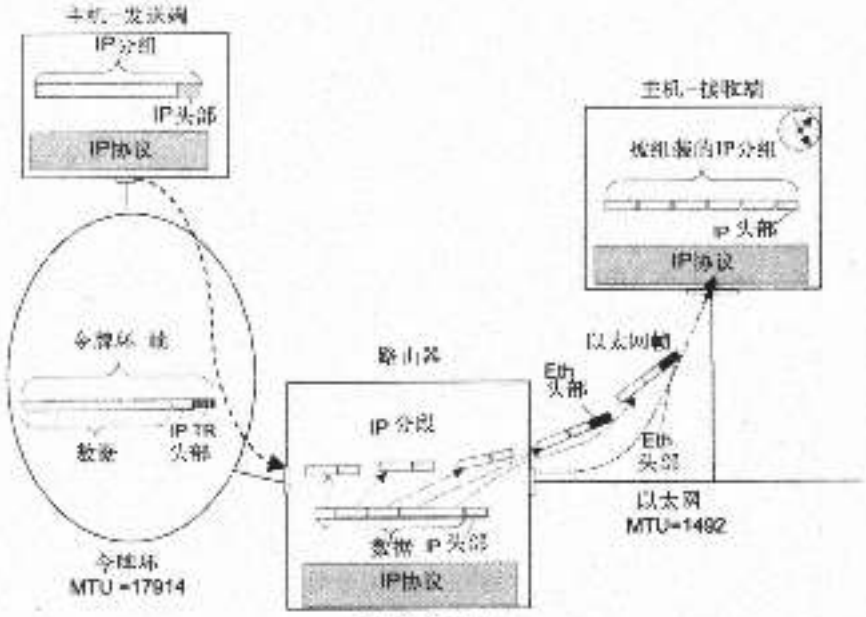


图18-19　分片

在帧路过路由器网络接口的网络层同时令牌环头部被剥离之后，路由器的IP实体将从分组中检索出目的网络地址。根据该地址，它判断出新抵达的IP分组必须被传送到MTU值为1 492的以太网。该值远远小于抵达输入接口的分组大小。因此，IP分组必须拆分成段。路由器从分组中检索出数据字段并将它拆分成下列大小的部分：四个各包含1 400个字符分段以及一个包含1 000字节的分片。注意，每个数据字段都是8的倍数。然后，IP实体形成新的IP分组，其中四个长度为1 400 + 20 = 1 420字节，最后一个分组为1 000 + 20 = 1 020字节。这些值均小于1 500字节，因此，这些分组符合以太网帧的数据字段要求。

最后，连接到以太网的目的主机接收到五个拥有共同标识符12456的IP分片。如果这些分片在一定的时间间隔内抵达，那么运行在目的主机上的IP就可以组装出源消息来。如果这些分组抵达的顺序与发送时不一致，则偏移量（offset）字段会指明它们组装时的正确顺序。

18.6　IPv6

自从20世纪90年代以来，TCP/IP栈即遇到了严重的问题。那时，因特网已经被广泛使用，大多数用于工业用途。比如，大多数组织开始使用因特网组建它们的公司范围网络作为传输系统，同时他们使用Web技术来访问公司信息。除此以外，因特网的电子商务被广泛应用，大众传媒以及娱乐企业也开始配备因特网，比如视频和音频广播以及交互式的游戏。

所有这些因素导致网络主机数目的急剧增长。在20世纪90年代早期，每隔30秒就有一台新的主机联入因特网。导致的结果是，流量类型改变了，QoS需求也变得异常急切起来。

18.6.1　TCP/IP栈的新方向

因特网委员会以及整个电信工业开始通过开发TCP/IP栈的新协议，如RSVP、IPSec、MPLS等，来解决新出现的问题。然而，即使这样，专家们也很清楚TCP/IP技术不能只通过增加新的协议来进

一步提高性能，有必要冒险进行栈核心，即IP，的现代化改造了。此方案的主要前提在于在不改变IP分组格式、重新加工IP头部字段的处理逻辑前提下，有些问题已经无法解决了。在这些问题中，最迫切的是可获取IP地址的短缺，该问题在不拓展源和目的地址字段大小时显然是无法解决的。

路由可延拓性成为批评界最常攻击的对象。关键是网络拥塞路由的快速增长，甚至现在仍然得处理包含几万条记录的路由表。同时需要记住的是路由表也执行一些辅助任务，比如拆分分组。针对此问题提出了一些解决方案，但是它们同样需要引入对IP的改动。

在为IP直接增加新功能的同时，还需要确保它与TCP/IP栈的新协议之间的密切交互。这同样需要往IP头部增加字段，处理这些字段必须由这些协议来完成。比如说，为了确保RSVP操作，需要为IP头部引入流量标志字段，而IPSec，则需要关于数据传输的特殊字段来保障其安全性的功能。

总之，因特网委员会已经决定改变IP，选择下列目标作为该现代化改造的主要目标。

- 创建一个可延拓的寻址方法
- 减少由路由器执行的操作
- 为传输服务质量提供保证
- 确保为网络上传输的数据提拱保护

对IP现代化领域的活跃研究以及与之关联的新协议的发展始于1992年，当时呈现在因特网委员会面前的有好几个下一代IP的不同版本：IPv7（由Ullman开发），TUBA（Callon），ENCAPS（Hinden），SIP（Deering）以及PIP（Francis）。

作为开发过程的结果，在1993年，ENCAPS、SIP和PIP等项目被合并到同一个提案的框架中，变为SIPP。在1994年7月，此提案被采纳为下一代IP发展的基础——**下一代因特网协议**（next generation Internet protocol）（IPng）。如今，IPv6缩写通常用以代表新一版本的IP。

登记IPv6文档的为RFC 1752，“对下一代IP协议的建议书”。命名为IPv6的基本协议集在1995年9月被IETF采纳。在1998年8月，定义公共IPv6架构标准的修订版（RFC 2460，“因特网协议，版本6说明（*Internet protocol, Version 6 Specification*）”）以及它独立的内容如寻址系统（RFC 2373，“IP版本6的寻址架构”）被采用。描述IPv6地址架构标准的最新版本（RFC 3513）最近被启用，应该是在2003年，被启用。

18.6.2 可延拓的寻址系统

更新的、第六版IP向IP网络寻址系统引入了重要的变化，这些变化与*地址位容量*（*address bit capacity*）的增长最为相关。

地址位容量（*address bit capacity*）。一个IPv6地址包含128位或16字节，这提供了为大量主机编号的可能性：340 282 366 920 938 463 463 374 607 431 762 211 456。这个数的范围如下解释：如果理论上可能的IP地址数分散在地球上所有居住者中（大约60亿人口），那么每个人将拥有的IP地址将多得无法想象，名义上为$5.7*10^{28}$个！

显然，地址长度如此显著地增长并不是为了减少地址短缺问题而在于提高TCP/IP栈操作的整体效率。

寻址功能如今的主要目标并不是地址空间的机械增长。与之相反，它旨在以引入几个新字段的代价来延伸其功能。

取代两层结构（网络号和主机号）的是，IPv6提供四层结构，其中三层用于标识网络，一层用于标识主机。由于地址层次结构增长的数目，新协议有效地支持了CIDR技术，CDIR技术、改进的组寻址系统和新的地址类型的引入，使新的IP版本有效地降低路由成本。

地址表示（*address presentation*）。该改变也包含纯粹是为了美观的效果，比如说，开发者建议用IP地址表示的十六进制形式取代其二进制形式。每四个十六进制数字被一个冒号隔开，比如，

典型的IPv6地址看起来如下：FEDC:0A98:0:0:0:0:7654:3210.

如果地址包含一长串0，该表示将被缩写。比如，前面提供的地址可以写成：FEDC:0A98::7654:3210

其中“::”省略号在一个地址中之可以使用一次。也可以省略地址中每个字段的起始0，比如，FEDC:0A98::7654:3210可以写成FEDC:A98::7654:3210.

同时支持两个IP版本——IPv4和IPv6——的网络允许为IPv4的4个低位字节使用二进制传统表示法。对于12个高位字节，则倾向于使用十六进制形式：0:0:0:0:FFFF:129.144.52.38或::FFFF:129.144.52.38.

地址类型（*address types*）。新版本IPv6，提供下列3种类型的地址：单播，多播，任播。地址类型由地址的几个高位值来定义，又叫做格式前缀（*FP*）。

- **单播**（unicast）类型的地址定义了终端节点或路由器单个接口的唯一标识符。此类地址的主要目的通常与IPv4中使用唯一地址的目的相一致，使用这类地址，协议将分组传送到目的节点的特定网络接口。然而，与IPv4比起来，IPv6中没有网络分类或固定划分网络号和主机号的8位界限的概念。单播类型的地址被分为多个子类型，分别反映当前网络中若干最常见的情况。
- **多播**（multicast）类型地址的主要目的是形成与IPv4组地址相类似的组。它的前缀为1111 1111格式并且识别通常关联到不同主机的接口组。携带此类地址的分组被送往拥有该地址的所有接口。多播类型地址同时用于IPv6取代**广播地址**（broadcast addresses），为达此目的，引入了一个特殊的组地址，它将子网中的所有接口联结起来。
- **任播**（anycast）类型地址是一种新型地址，它与多播类型相似，指定一个接口组。然而，携带此类地址的分组只被送往组中的一个接口。原则上，这是根据路由协议度量出的“最近”的一个接口。语法上，一个任播地址与一个单播地址并无差别且与单播地址处于相同的地址范围。任播类型的地址只能被分配给路由器接口，属于同一任播组的路由器接口有着自己的单播地址和共同的任播地址。当发送端用指定所有中继路由器的IP地址的方法定义这个分组的路由时，此类型的地址被定位到源路由。比如，提供商可以分配相同的任播地址给所有的路由器并且指定该地址给客户。如果客户想要分组从他的网络传输途经属于该ISP的网络时，将此任播地址指定在源路由地址链中就足够了。分组就会经最近的路由器传输给那个提供商。
- 在IP的第六版本中，存在着自治网络中使用的**私有地址**（private address）。与IPv4相比较，在IPv6中的这些地址有着特殊的格式。IPv6有两种本地使用的地址：
- 首先，存在着不被分割成子网的网络地址（并且不使用路由）。它们叫做**链路-本地地址**（link-local addresses）并且有着下列格式的10位前缀：1111 1110 10。链路-本地地址只包含接口标识符的64位字段；除了FP以外所有其他位均置为0，既然这种情况下不需要使用子网号。
- 其次，此组中含有用于将网络分割成子网的本地地址。此类地址叫做**站点-本地地址**（site-local addresses）；它们有着下列格式的前缀：1111 1110 11，并且与链路-本地地址相比，它包含额外的2个字节字段用于存储子网号。

单播地址的主要子类型是全局聚合唯一地址。这种地址可以被聚合以简化路由。与IPv4版本的唯一地址相比较，IPv4包含两个字段——网络号和主机号——IPv6的全局聚合唯一地址结构更加复杂，包含六个字段（图18-20）。

- 此类地址的*FP*包含3位，值为001。

下列三个字段——顶层聚合（top-level aggregation）（TLA），下一层聚合（next-level aggregation）（NLA）以及站点层聚合（site-level aggregation）（SLA）——代表着网络标识的三层。

- *TLA*旨在标识最大提供商的网络，这个前缀值代表着提供商所拥有地址的共同部分。分配给此字段相对较少的数目，13，是特意为限制顶层因特网的主干路由器中路由表的大小而选取

的。此字段允许顶层提供商的8 196个网络。这意味着描述这些网络间路由的记录数同时被限定为8 196，这将加速主干路由器的操作。紧邻的8位被保留以备日后使用，需要时可以拓展TLA字段。

3位	13位	8位	24位	16位	64位
格式前缀 (FP)	顶层聚合 (TLA)		下一层聚合 (NLA)	站点层聚合 (SLA)	接口标识符

图18-20 IPv6分组中全局聚合唯一地址的结构

- 下一层的前缀*NLA*，旨在为中小型提供商的网络计数。*NLA*字段的大小允许通过地址聚合创建多层地址层，此结构反应出ISP的多层结构。
- *SLA*用于单个注册客户的寻址网络，比如代表同一公司范围网络部分的子网。假定提供商指派给一个特定的公司网络号中含有特定的TLA和NLA字段值，组合起来可以模拟IPv4网络号。地址的剩余部分——SLA和接口标识符字段——由管理员配置。公司网络管理员控制着地址形成过程且不会与提供商的步骤相一致。同时，接口标识符（*interface ID*）字段有着特定的目标——它必须存储主机的物理地址。在这一层上，同样可以将小型子网的地址聚合成大型子网。SLA字段充分地提供了便利组建公司特定地址结构的可能。
- 接口标识符（*interface ID*）是对IPv4主机号的一个模拟。IPv6中的区别在于，通常情况下，接口标识符仅仅匹配它的本地（硬件）地址，而不代表着管理员任意分配给它的主机号。接口标识符长64位，可以为一个MAC地址（48位）、一个X.25地址（长至60位）、一个ATM终端节点地址（48位）或者一个ATM虚拟连接号（长至28位）。理论上，它应该可以提供使用所采用技术的本地地址的可能性。这种IPX风格的解决方法使得不再需要利用ARP，因为IP地址映射到本地地址的过程变得很简单——它被简化到只需要丢弃地址最高位部分。而且多数情况下，不再需要手动配置终端节点。这是因为主机从硬件（网络适配器等）得到地址的低位部分，接口标识符，同时路由器负责通知终端节点地址的高位部分，网络号。

很显然，IPv6有了如此大量的网络，子网化（即，利用掩码将网络拆成子网）的操作便失去了它的价值。另外一方面，相反操作——超网化——变得格外重要。IPv6标准的开发者将地址聚合视为有效利用IP新版本中地址空间的主要手段。

示例 假定客户端从ISP获取了一个IPv6地址池，定义为：20:0A:00:C9:74:05/48。

分析一下这个号码。因为其前3位被设为001，是全局聚合唯一地址。使用此类地址的标准形式表示这个前缀20:0A:00:C9:74:05/48（图18-21）。

这个地址属于顶级提供商，所有子网有着共同的前缀20:0A/16。此提供商可以分配相同前缀的地址范围给二级提供商。这个共同的前缀将建在顶级提供商前缀以及部分NLA字段的基础上。分配给前缀的NLA字段长度由顶级提供商的掩码决定。假设示例中掩码的高位包含32个1，那么二级提供商的前缀如下所示：20:0A:00:C9/32。

因此，二级提供商有16位NLA字段来为客户的网络编号。二级提供商的客户列表可以包含三级或更小的提供商以及注册用户——公司和组织。假设NLA字段的下一个字节（01110100）被提供商用以传递给三级提供商，后者反过来，使用NLA字段的最后字节为客户指定地址池。因此，提供商的三层结构就形成了20:0A:00:C9:74:05/48前缀来分配给客户端。

IPv6给客户端留下2个字节以供配置网络号（*SLA*字段），8个字节以供配置主机号（接口标识符（*interface ID*）字段）。

有如此大量的子网号，管理员可以用于多种用途。比如说，他/她可以选择网络平面寻址从

65 535个地址范围中选择特定值，而不使用保留的地址。在大型网络中，基于地址聚合的层次网络结构可以成为组织地址空间更加有效的方法，因为它减少了公司路由器上路由表的大小。这种情况下，用到了传统的CIDR技术，然而，这是由公司网络管理员而不是ISP来实现的。

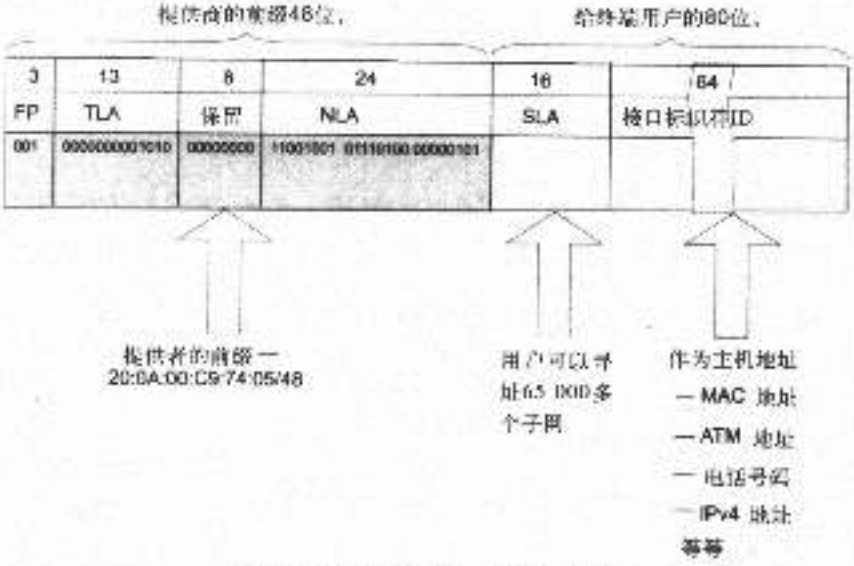


图18-21　全局聚合唯一地址的示例

除了刚才提到的全局聚合唯一地址，还存在着其他类型的单播地址，

- IPv6中的回环地址0:0:0:0:0:0:0:1与IPv4中的127.0.0.1功能相同。
- 未定义地址::，由全0组成，与IPv4中的0.0.0.0相仿。这个值表明主机缺少分配的IP地址。该值不会出现在IP分组的目的地址中，如果它出现在源地址中，这意味着在主机获知IP地址之前分组就已经发送出来了（比如，在从DHCP服务器得到地址之前）。

通常认为在IPv6基础上工作的因特网分段会与其他使用IPv4的因特网共存一段时间。为了确保支持IPv6的主机能够自动使用将IPv6分组通过IPv4网络传输的技术，发展了一种特殊的地址子类型。这种子类型的地址在IPv6地址的低四位字节中携带IPv4地址，12个高位字节都用0填充（图18-22）。这种单播地址类型简化了IPv6地址到IPv4地址的转化步骤，被命名为*IPv4-兼容的*（*IPv4-compatible*）的IPv6地址。

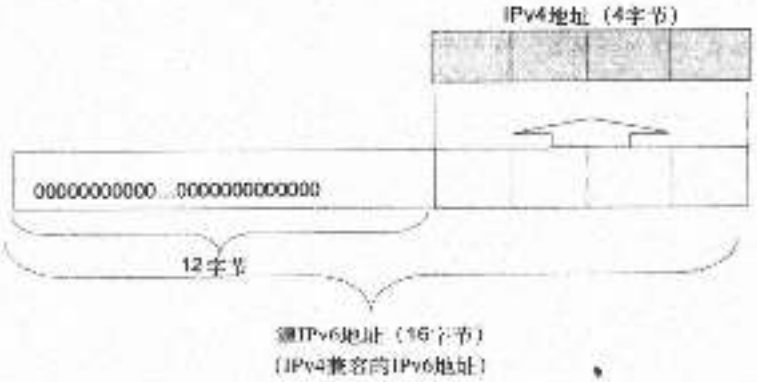


图18-22　IPv6到IPv4的转换

还有一种携带IPv4地址的IPv6地址——IPv4映射的IPv6地址，这种地址是为了解决相反的问题——将IPv4分组在根据IPv6分段操作的因特网中传输。这种类型的地址第四个字节包含一个IPv4地址。它们的高十个字节用0填充，第五和第六个字节包含全1，以表明该主机支持IPv4（图18-23）。

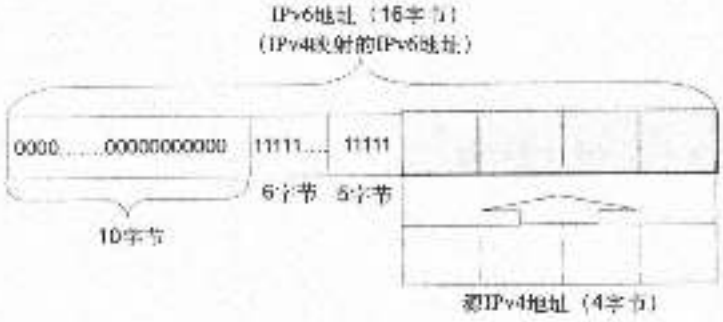


图18-23 IPv4到IPv6的转换

关于IPv6子类型的字段研究还未完成。IPv6地址空间中只有15%有着目标定义明确，地址的其余部分还等着为解决众多网络问题作出自己的贡献呢。

18.6.3 灵活的头格式

改变IPv6头部格式的一个重要目标是减少超负荷——也就是，减少随每个分组一起传送的控制信息量。为了达到这个目的，在新版本IP中引入了*主要*（*main*）头部和*附加*（*additional*）头部的概念。主要头部一直存在，而附加头部（additional header）则是可选的。举个例子，附加头部可能包含被分组的分段信息，使用源路由技术时分组的完整路由以及传输数据的安全信息等。

主要头部的固定长度为40字节，其格式在图18-24中说明。



图18-24 主要头部的格式

下一个头部字段对应于IPv4的协议字段，定义了紧随着当前分组的头部类型。每下一个附加头部也有下一个头部字段。如果一个IP分组不含有附加头部，那么该字段将包含指派给这种协议的值，比如TCP、UDP、RIP、OSPF或其他由IPv4标准定义的协议。

目前，在IPv6标准提议中提到了下列类型的附加头部：

- **路由**（routing）——用于在使用源路由时指定完整路由
- **分片**（fragmentation）——包含与IP分组分片相关的信息。该字段在终端节点上处理。
- **认证**（authentication）——包含终端节点认证以及确保IP分组的完整性所需的信息
- **封装**（encapsulation）——包含利用加密和解密来确保传输数据机密性所需要的信息
- **逐跳选项**（hop-by-hop option）——设置根据逐跳算法处理分组时使用的特殊参数
- **目的选项**（destination option）——包含目的节点的辅助信息

因此，IP分组格式看上去如图18-25所示。

因为分组路由只需要主要头部（实际上所有附加分组只在终端节点上处理的），这就减少了路

由器上的负担。另一方面，使用大量附加参数的可能性拓展了*IP*功能并且使得它易于引入新机制(*extends the IP functionality and makes it open for introducing new mechanisms*)。

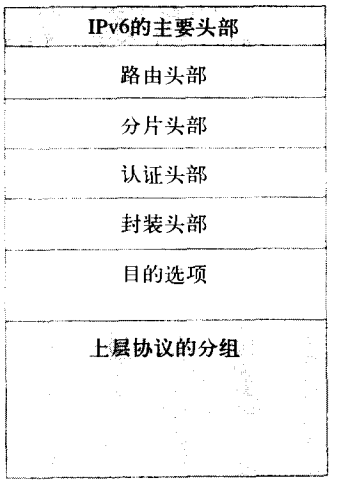


图18-25　IPv6分组的结构

18.6.4　减少路由器的负荷

为了提高因特网路由器在主要功能、分组传递方面的性能，IPv6提供了将路由器从执行辅助任务中解脱出来的方法：

- 将分片功能从路由器转移到终端节点。在IPv6中，终端节点必须判定连接源主机到目的主机路由沿途中的最小MTU（该技术被称为MTU路径查询，曾被用于IPv4中）。IPv6路由器不执行分组分片。相反，它们仅仅发送ICMP消息**超长分组**（packet too long）给终端节点。收到这种消息，主机必须降低分组大小。
- 地址聚合，减少了路由器上地址表的大小。表查询时间和需要修改所需的时间也因此减少了。与此同时，路由协议创建的辅助流量也减少了。
- 广泛使用源路由，其中源节点指定了分组在网络中传输的完整路由。该技术将路由器从为选取下一个路由器而搜索地址表的重担中解脱出来。
- 放弃处理可选的头部参数。
- 将节点的MAC地址作为网络号，使得路由器不再需要使用ARP。

IP的新版本，IPv6项目的一个重要组成部分，提供了内嵌的数据保护工具。在网络层实施保护工具将使它们对于应用程序不可视，因为在IP层和应用程序之间总是存在着传输层协议。使用这个解决方法时，将不需要重新设计应用程序。内嵌安全工具的IP新版本称为IPSec，此协议的细则将在第24章中提到。

从IPv4到IPv6的迁移才刚刚开始，因特网段存在于两个版本都支持的路由器。这些段利用因特网彼此相连，因此形成**6骨干**（6Bone）。

小结

- IP在因特网节点间解决了数据传送的问题。既然是一个数据报协议，它不能保证数据传送的可靠性。
- IP的特性，与其他网络协议不同（比如IPX），它在MTU不同的网间传送分组时，执行其动态分段的能力。
- IP分组的最大长度为65 535字节。头部长度通常为20个字节，包含发送端和接收端网络地址、分片参数、分组TTL、校验总和以及其他数据的信息。
- IP路由表的格式依赖于路由器的特定实现。虽然显示在屏幕上表的形式不一样，但所有的表都包含两个必要字段：源地址和目标地址，否则路由就无法执行。
- 来自于三个源的记录，应用于路由表中。第一个来源，TCP/IP栈软件，作为配置结果，将连接网络和默认路由器以及关于特殊用途地址的记录保存在表记录中。第二个来源，管理员可以手动输入特定路由或默认路由器的记录。最后，路由协议自动将现存路由的动态记录保存入表中。

- 掩码提供了重叠IP网络的有效方法，它将单个网络分割成多个子网（子网化）或将多个网络聚合成一个大型网络（超网化）。
- 无类别域间路由技术（CIDR）在未来的IP网络中扮演着重要角色。该技术解决了两个重要问题。首先是确保了可用地址空间的有效利用；其次是减少了路由表的记录数，因为一条记录可以代表着拥有共同前缀的多个网络。
- 自20世纪90年代以来，TCP/IP就面临着日益严重的问题。不改变IP分组的格式以及IP头部字段的处理逻辑将不能解决问题。因此，因特网委员会决定开发IP的新版本，IPv6。这个革新有着如下目标：创建可扩展的寻址系统，通过减少路由器执行的操作来提高网络带宽，确保传输服务质量，确保数据网络传输的安全性。

复习题

1. 如何证明IP的不可靠性？
2. 对比网桥或交换机与路由器的地址表。描述这些表是如何创建的，它们包含什么信息？表的大小受哪些因素影响？
3. 考虑一个主干因特网路由器，下面哪些记录存在于含有目的地址字段的路由表中：
   A. 所有因特网的网络号
   B. 部分因特网（B）的网络号
   C. 部分网络号以及某些定义了特殊路由的因特网主机的完整地址
   D. 特殊目的地址，如127.0.0.0或255.255.255.255（D）
4. 路由表中可包含多少条默认路由的记录？
5. 提供几个原则什么情况下需要使用特定路由的例子。
6. 当路由是基于掩码执行时，IP分组包含掩码吗？
7. CIDR技术提供了什么好处？是什么因素阻止了它的广泛使用？
8. 连续IP地址池的前缀长度与该池中包含的地址数存在着什么联系？
9. 为什么默认路由的记录总是包含掩码为0.0.0.0的目的网络地址0.0.0.0？
10. 下列网络成分中哪个可以执行分片：
    A. 只有计算机
    B. 只有路由器
    C. 计算机、路由器、网桥和转换器
    D. 计算机和路由器
11. 当分组在传输过程中被分片并且超时后其中一个分片仍然未到达目的主机时会发生什么？
    A. 发送端主机的IP实体将重新发送丢失的分片
    B. 发送端主机的IP实体将重新发送整个分组，其中包含丢失的分片
    C. 目的主机的IP实体将丢弃收到的该分组所有的分片。发送端主机的IP实体不会采取有关下重新传输此分组的任何操作。
12. 图18-26显示了一个配备两台网络适配器的计算机，两个适配器各连接一个网络块。此计算机上运行Windows 2000，请问计算机A可以与另外一个网段中的计算机B交换数据吗？
13. 这些网段使用不同类型链路层协议，比如以太网和令牌环。这会影响到前一问题的答案吗？



图18-26 连接于一台计算机的两个网段

14. IPv6的管理员如何使用掩码？
    A. 忽略，因为不需要使用掩码
    B. 利用超网化
    C. 利用子网化
    D. 利用超网化以及子网化
15. 如果有人说广播是多播的一个特殊应用，他说得对吗？如果说广播是任播的一个特殊应用，是否正确呢？
16. 同一网络接口同时拥有多种IPv6地址：单播、任播、多播、这可能吗？

练习题

本章的18.4.4节包含了一个网络规划的示例。在此例中，网络管理员设计了如下需求：600个以太网地址，200个令牌环网地址，10个DMZ地址以及4个用于网络连接的地址。请试图解决在300个工作地点计划用令牌环网时的同一问题。

- 必须从ISP获取怎样的地址池？在此例中，ISP将提供给客户连续的地址池。
- 管理员将如何分配地址给这四个网络？
- 路由器R1和R2的路由表看起来如何？

第19章 TCP/IP栈的核心协议

19.1 引言

本章我们从研究TCP和UDP开始，这两个协议在应用程序和网络传输基础设施之间扮演着中间人的角色。尽管因特网协议（IP）所相关的因特网层的主要目标，是在网络接口之间进行数据传输，但由TCP和UDP负责执行的传输层的主要任务，包含了运行在网络上的计算机上的**应用程序进程**（Application process）之间的数据传输。

然后，我们考虑一下旨在自动构建路由表的路由协议，这是转发网络层分组的基础。与IP或IPX等网络协议相比，路由协议是可选的，因为路由表可由网络管理员手动创建。然而，在有着复杂拓扑和诸多可选路由的大型网络中，路由协议负责自动化构建路由表过程的重要工作。在网络结构变化时，比如网络失败或引入新路由器和通信链路时，它们同样可以发现新路由。

我们同样会涉及到因特网控制信息协议（ICMP），它是通知发送端有关它的分组未被送到目的地的工具。除了用于诊断外，ICMP还可用于**网络监控**（*network monitoring*）。比如，ICMP报文被用于常用IP网络监控工具的基础，如流行的ping和traceroute。

19.2 TCP和UDP运输层协议

正如前文所述，传输层的主要任务由传输控制协议（*TCP*）完成，它定义于RFC 793以及用户数据报协议（*UDP*），它定义于RFC 768。该任务包含运行在联网计算机的应用程序之间的数据传输。既然TCP和UDP与同一层相关，所以它们有着很多的共同点。两个协议都为上一层协议即应用程序协议，提供接口，它将进入主机的数据传输给合适的应用程序。与此同时，两个协议都使用**端口**（port）和**套接字**（socket）的概念。两个都通过将它们的PDU封装进IP分组来支持下层的接口，网络IP层。与应用层协议相似，TCP和UDP的协议实体都只安装在终端节点上。然而，正如你在后文中会看到的，TCP和UDP之间的差别比它们的共同之处还要多。

19.2.1 端口

每个计算机可以执行多个进程；而且，每个应用程序进程可以拥有多个作为数据分组目的地址的访问点。因此，分组通过IP传送到目的主机的网络接口之后，还需要将该数据传送到某一特定进程。

同时必须执行相反的任务：运行在同一终端节点上的不同应用程序接收到的分组由相同的IP处理。因此，协议栈必须提供从不同应用程序“收集”分组并传递到IP的方式。TCP和UDP都可以完成此项工作。

由TCP/UDP执行的从多个应用程序服务接收数据的过程叫做**多路复用**（multiplexing），与其相反的过程，TCP/UDP用来将来自网络层的分组在更高层服务集中进行分配的过程，叫做**解多路复用**（demultiplexing）（图19-1）。

对于每个应用程序端口，TCP和UDP维护着两个队列：从网络到达这个应用程序的分组队列以及应用程序发送给网络的分组队列。到达运输层的分组由操作系统组作织作为服务于不同应用程序进程的不同访问点的队列集。在TCP/IP术语中，这类系统队列叫做**端口**（port）。（不要将应用程序端口与硬件端口相混淆，后者是网络设备的网络接口。）注意，同一应用程序的输入和输出队

到被认为是同一个端口。对于唯一的和确定标识的端口，会为它们分配端口号，端口号用于对应用程序寻址。

为应用程序分配端口号存在着两种方法：集中式（*centralized*）和本地式（*local*）。每种方法都有它自己的端口号范围：对于集中式方法，分配范围是0到1 023，而本地方法则使用1 023到65 535范围内的端口号。

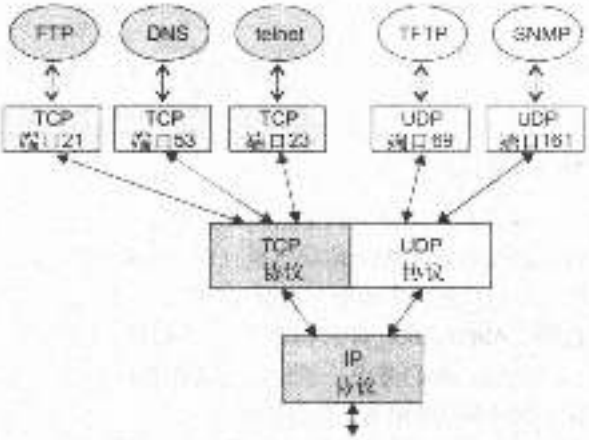


图19-1　运输层的多路复用与解多路复用

如果进程为常用的公共服务，比如文件传输协议（FTP）、telnet、HTTP、TFTP或DNS，它们被分配标准端口号（standard port numbers），又叫做知名端口号（well-known port numbers）。这些号都列在因特网标准中——RFC 1700和RFC 3232。比如，21号分配给了FTP服务，23号分配给了telnet服务。分配的地址是唯一的，这意味着任何其他应用程序都不允许使用。

对于那些没有如此大范围流行和使用的应用程序，端口号由应用程序开发者或操作系统以相应应用程序的请求分配出去。在每台计算机上，操作系统维护着已分配和空闲的端口号列表。当运行在本地计算机上的某一应用程序请求到达时，操作系统会为之分配首先可用的端口号。这类号被称做动态的。接下来，所有网络应用程序都将使用分配给它的端口号来访问此应用程序。在应用程序中断执行之后，分配给它的本地端口号将返回给可用端口号列表并可供其他应用程序使用。**动态端口**（dynamic port）号在每台计算机范围内是唯一的，然而，运行在不同计算机上的应用程序拥有相同端口号是很常见的现象。原则上，知名应用程序（DNS、WWW、FTP、telnet、等等）的客户端部分从操作系统中接收动态端口号。

这些与端口相关的信息同样适用于传输层的两个协议。通常，利用TCP为应用程序分配端口号与用UDP的同一操作之间并不存在着相互依赖关系。利用UDP将数据传到IP层的应用程序获取的号叫做UDP端口（UDP port）。同样，利用TCP分配给应用程序是TCP端口（TCP port）。

这两种情况下，都可能是指定或动态端口号。TCP和UDP端口的分配范围正好一致：与0到1 023用于指定端口号，1 024到65 535用于动态端口号。然而，指定给TCP和UDP端口号之间并不存在着联系。即使TCP与UDP端口号重合了，它们分别标识着不同的应用程序。比如，一个应用程序可能被分配TCP端口1 750，而另外一个应用程序使用UDP端口1 750。当某个应用程序可以选择TCP和UDP时，它被分配一个重合的TCP和UDP端口号以便记忆。打个比方说，DNS可以使用TCP端口53或者UDP端口53。

19.2.2　UDP

UDP是一个数据报协议（即以数尽力服务（best efforts）原则工作的协议，它并不建立逻辑连接），作为一个数据报协议，UDP并不保证消息的传递。层面也不为IP可靠性的不足做出补偿，IP也是一个数据报协议。

UDP的数据单元叫做UDP分组（UDP packet）或**用户数据报**（user datagram），每个UDP分组携带一个独立的用户消息（图19-2），这导致了一个天然的限制：一个UDP数据报不能超过IP数据字段的长度，后者反过来，受限于下层网络技术的帧大小。因此，一旦UDP缓冲超长，该应用程序数据就被丢弃。UDP头部包含4个2字节字段，包含发送端和接收端的端口号、校验和以及数据

报长度。

这里提供了一个填充了值的UDP头部分段：

源端口 = 0x0035

目的端口 = 0x0411

总长度 = 132（0x84）字节

校验和 = 0x5333

在此UDP数据报中，数据字段紧随着头部之后，长度为132-8字节。它包含来自DNS服务器的消息，这可以通过源端口号来判断，源端口 = 0x0035，对应着DNS服务器的标准端口号，53。

从头部的简单性看出，UDP并不是一个复杂的协议。它的功能被缩减到网络与应用层之间的多路复用和解多路复用。考虑一下UDP是如何解决解多路复用问题的，利用端口号看起来是执行此任务很自然的一个途径。携带UDP数据报的帧到达主机的网络接口，被栈协议顺序处理，最后，传给UDP。UDP从分组头部检索出目的端口号并将数据传送到相应的应用程序的合适端口（即，执行解多路复用任务）。

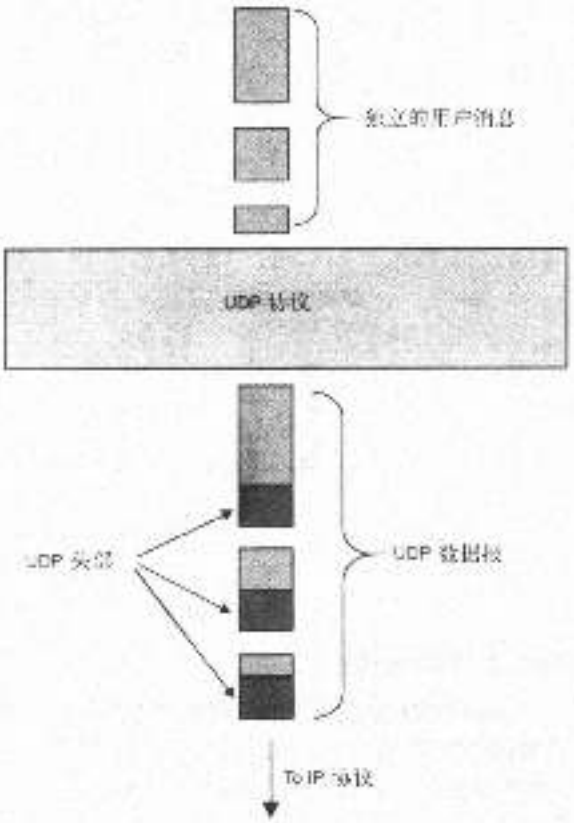


图19-2 形成UDP数据报

这个解决方案看起来既符合逻辑又简单。然而实际上，当同一个应用程序的多份拷贝运行在同一台终端节点上时，它并不奏效。假设两台DNS服务器运行在同一主机上，两个都使用UDP传递消息（图19-3）。此DNS服务器被分配了一个知名UDP端口，53。同时，每台DNS服务器可能都为它自己的客户端服务，有着自己的数据库，并有着独立的定制设置。当DNS客户端的请求到达该计算机的网络接口时，UDP数据报会指定端口号53，这将关联到两台DNS服务器。那么UDP该将请求递交给哪台服务器呢？为了避免不确定性，使用下面解决方法：为安装在同一计算机上同一应用程序的不同拷贝分配不同的IP地址。此例中，DNS服务器1有地址IP_1，而DNS服务器2则被分配IP_2地址。

因此，网络中的应用程序（甚至是在一台计算机中）的地址由IP地址和UDP端口号对来明确定义。这个对叫做**UDP套接字**（UDP socket）。套接字的使用使得UDP正确解多路复用成为可能。

说明 在这个关系中，我们必须澄清根据分组在协议栈传输方式的简单化模式。正如前面章节中提及的，在IP处理过网络分组之后，此分组的头部就被丢弃了，只有分组的数据字段内容被继续传递——这可能是UDP数据报。然而，在解释这个机制时，我们忽略了一个重要细节：与数据字段一起，从分组头部检索出的目的IP地址被传递给传输层。UDP从UDP数据报头部检索出端口号，在套接字（目的IP地址以及目的端口号）的基础上，执行解**多路复用**（demultiplexing）。

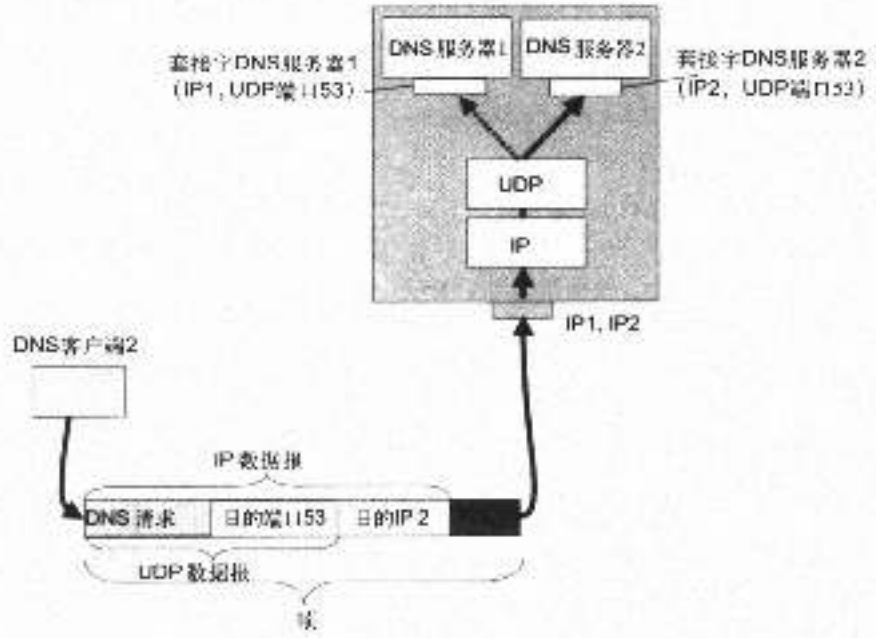


图19-3　基于套接字的UDP解多路复用

19.2.3　TCP段格式

高层协议提供给TCP的信息被TCP认为是**无结构字节流**（unstructured stream of byte）。到达的数据由TCP暂时保存。该协议然后"切除"一些连续的数据段[⊖]，为它提供头部，并传递到网络层（图19-4）。

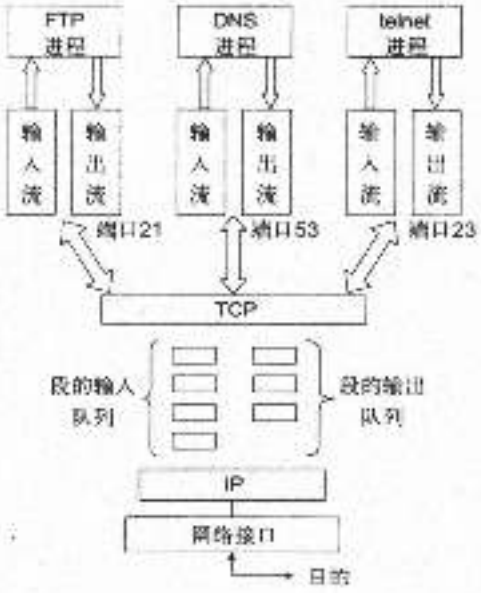


图19-4　从一个无结构字节流形成TCP分段

TCP段的头部字段远多于UDP头部的字段，这是因为TCP拥有更加高级的能力：

- **源端口**（source port）——该字段占用2字节，标识发送端进程。
- **目的端口**（destination port）——该字段占用2字节，标识目的进程。
- **序列号**（sequence number）——该字段占用4字节，标识定义了相对于发送数据流的段偏移量的字节号（即，段中第一个字节的编号）。
- **确认号**（acknowledgment number）——该字段占用4字节，值为所收到段的最大字节号再加上1，该值在确认时使用。如果ACK检查位被设置，那么该字段包含数据报发送端想要接收的队列中的下一个值。
- **头部长度**（Hlen）（header length）——这个4字节字段指出TCP段的头部长度，以32位字来衡量。头部长度并不固定并可以根据选项字段中设置的参数而改变。
- **保留**（reserved）——这个保留字段占用6位，它被保留以供将来使用。

⊖　注意术语段既可以表示作为一个整体的数据传输单元（数据字段和TCP头部），也可以仅仅代表数据字段。

- 编码位（code bits）——该字段占用6位，包含本段类型的辅助信息。此信息通过设置下列字段位来申明：
  - 紧急数据（urgent data）——这是一条紧急信息。
  - ACK——确认已收到段。
  - PSH——请求在缓冲区尚未填满时发送消息。注意TCP可以等待缓冲区填满之后再发送段。如果需要立即传送，应用程序必须利用push参数来通知协议。
  - RST——申请重置连接。
  - SYN——此消息用于在建立连接时同步传输数据的计时器。
  - FIN——该段落中明发送端已经发送出被传输数据流的最后一个字节。
- 窗口（window）——该2字节字段申明了数据字节数，从确认号中指明的字节号开始，当前段的发送端一直等待着该字节的到达。
- 校验和（checksum）——该2字节字段包含着校验和。
- 紧急指针（urgent pointer）——该字段占用2字节，与*URG*代码位一起使用并且申明即使存在着缓冲区溢出也必须紧急接收的数据末端。例如，如果有些数据需要不按顺序被送往目的应用程序，那么发送端应用程序必须利用紧急数据参数通知TCP。
- 选项（option）——该字段为变长并且可以省略，它的最大长度为3字节，用于解决一些辅助任务。比如，选择最大段长。选项可以位于TCP头部的末端，其长度必须是8位的倍数。
- 填充（padding）——该字段长度不固定，这是个用于补充头部字段使得它的长度为32位字的整数倍的一个伪字段。

19.2.4 作为TCP可靠性基础的逻辑连接

TCP和UDP两上层区别在于TCP必须执行一些额外的任务。此任务包括确保消息在网络上的可靠传输（*reliable delivery*），其所有节点都使用消息传递的不可靠IP数据报协议。

图19-5显示了由安装了IP实体的路由器互连接的网络。安装在终端节点上的TCP实体通过彼此之间（*one another*）建立**逻辑连接**[⊖]（logical connections）来解决确保数据可靠交换的问题。TCP确保传输的分段不会丢失、重复或者乱序抵达接收端。

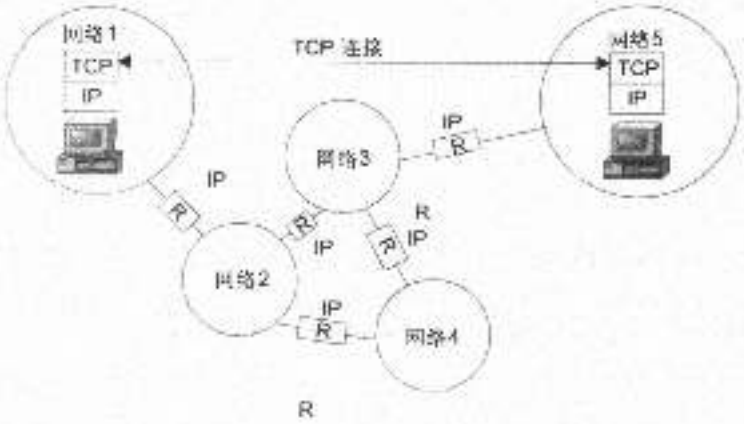


图19-5 TCP连接在终端节点之间创建了一个可靠的通信信道

- 在建立逻辑连接时，TCP实体协商数据交换过程的参数。在TCP中，每个参加者发送下列参

⊖ 参见第3章的3.3.4节。

数给它的合作伙伴：它准备好可以接收的段最大长度（*maximum size of segment*）。

- 它所能允许另外一个参加者往它的方向传输的数据最大容量（*maximum volume of data*）（可能是好几个段），即使这个参加者尚未收到前一拨发送数据的确认（该参数叫做**窗口大小**(window size)）。
- 字节的起始号（*starting number of the byte*），从此号开始在当前连接的帧中对数据流进行计数。

作为在TCP实体双方中进行协商的结果，定义了连接参数。有些参数在整个连接中保持不变；有些参数则会被调整。比如说，发送端的窗口大小根据接收端的缓存负荷以及网络操作的整体可靠性而动态改变。创建连接同样意味着每一台计算机的操作系统为组织缓冲区、定时器和计时器分配了资源。这些资源自创建时候起即被连接占用，直到连接中断才得以释放。

TCP逻辑连接由套接字对（*a pair of sockets*）来唯一标识。

每个套接字可以同时加入多个连接，比如说，假设有三个应用程序的三个套接字：(IP_1，n_1)，(IP_2，n_2) 和 (IP_3，n_3)。IP_1，IP_2和IP_3是它们的IP地址，n_1，n_2，n_3是它们的TCP端口号。在这种情况下，可以创建下面的连接：

连接1——{ (IP_2，n_2)，(IP_1，n_1) }

连接2——{ (IP_1，n_1)，(IP_3，n_3) }

连接3——{ (IP_2，n_2)，(IP_3，n_3) }

图19-6显示了套接字 (IP_1，n_1) 创建的连接1和3。

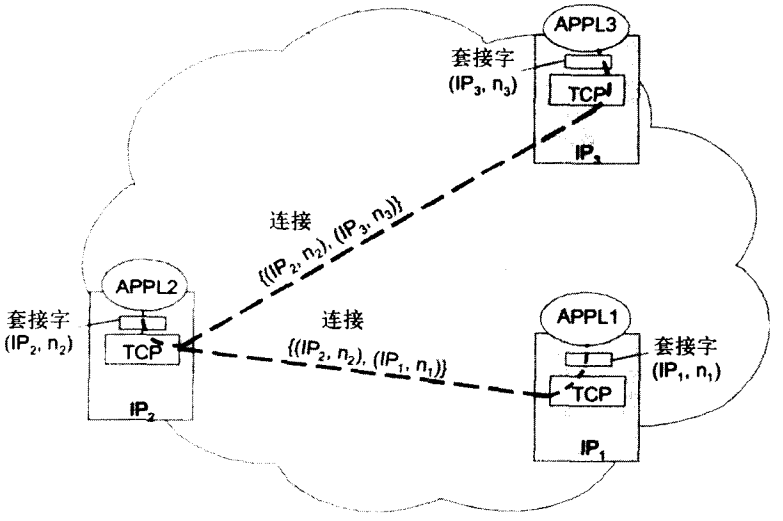


图19-6　一个套接字可以参与多个连接

现在，让我们来解释TCP是如何执行解多路复用任务的。考虑下面的例子：假设某一ISP提供网络宿主服务，意味着客户端可以在该ISP的服务器上安装它们的网络服务器。这个网络服务器基于HTTP应用层协议，即使用TCP。TCP通过监听知名端口80来等待网络客户端（浏览器）的查询。

图19-7显示了宿主了两台网络服务器的情况——www1.model.com，IP地址为IP_1，以及www2.tour.com，地址为IP_2。每台服务器可以同时服务多个客户端，且这些客户端也可以同时与WWW1和WWW2一起工作。每个客户端的工作需要服务器稳定地保存浏览页面数、连接参数等等——也就是，创建一个独立的逻辑连接。TCP为每个客户端-服务器对创建这种连接，而每个连接也是由相应的套接字对来唯一标识。

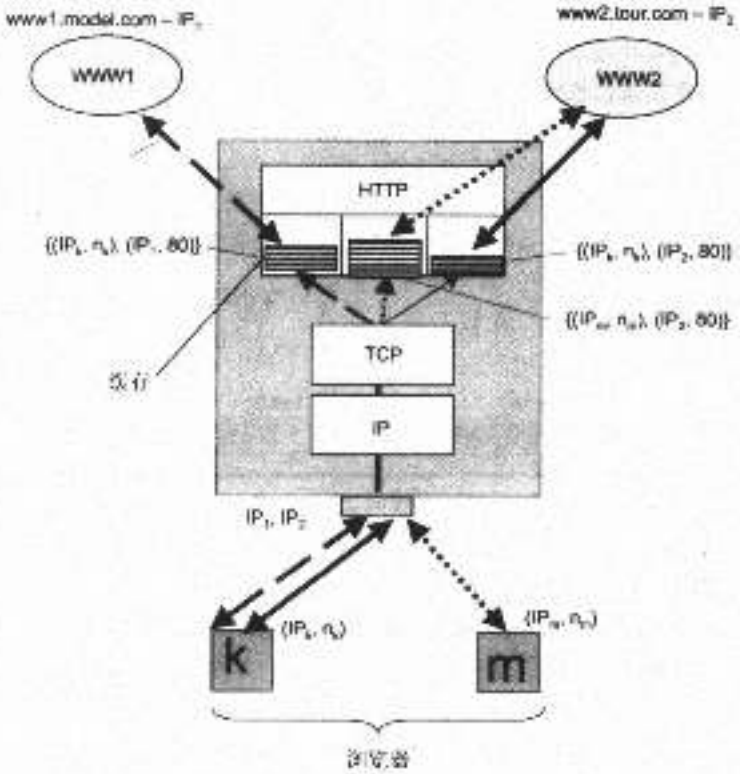


图19-7 基于TCP连接的解多路复用

图19-7显示了套接字为（IP_k，n_k）和（IP_m，n_m）的两个浏览器，浏览器k的用户同时访问WWW1和WWW2两个服务器，与其中与每台服务器相关的单独连接的存在确保了信息流分离，用户从不需要询问是哪台服务器发送来哪个页面。与浏览器k的用户同步，浏览器m的用户访问了服务器WWW2。这种情况下，两个用户都在单个逻辑连接范围内工作，这很好地分离了这些用户的信息流。图19-7同时还显示了多个缓冲区，其个数由逻辑连接数而不是网络服务器或客户端的数目来定义。根据发送端和接收端套接字的值来决定发送给这些缓冲区的消息。

因此，TCP是基于逻辑连接的基础上执行解多路复用的（*carries out demultiplexing on the basis of logical connection*）。

19.2.5 序列号和确认号

在TCP中，需要确认已建立连接框架中的每个段是否被正确传输。确认是确保可靠通信的一个传统手段。TCP使用的是一种特殊的确认机制，滑动窗口算法[⊖]（sliding window algorithm）。

TCP中实现的滑动窗口算法有一个特殊的特性：尽管每次传输的数据单元为段（*segment*），但窗口被定义为来自上一层的无结构数据流的有序字节集并被TCP缓存。

在建立连接时，双方协商字节的起始号，从这里开始在连接期间将一直执行计数。每一方都有它自己的初始号，每个段的标识为它的第一个字节号。段范围内的字节也有编号，这样紧跟着头部的数据的第一个字节拥有最小的编号并且其后的字节编号逐步递增（图19-8）。

当发送端发送了一个TCP段时，它将自己第一个字节的编号存入头部的序列号字段。因此，

⊖ 更多细节，请参见第6章的6.4.4节。

在图19-9中，接下来的号码被用作段标识：32 600，34 060，35 520等等。基于这些编号，接收端的TCP实体将特定段与其他段区分开来并且将接收到的段放在公共字节流中。除此之外，它还能够判断出该段是否是重复或已收到的段，是否在两个已接收段之间存在着丢失段等。

段的接收端发送一个确认信息作为响应。确认消息是一个段，接收端将接收段的最大字节号递增1之后存入其中，此号叫做确认号（*acknowledgment number*）。对于图19-9中所示的段，每个段的最后字节号递增1后作为接收确认（一个确认号）使用。对于第一个发送的段，这将是号34 060，对于第二个段，这将是号35 520，依此类推。

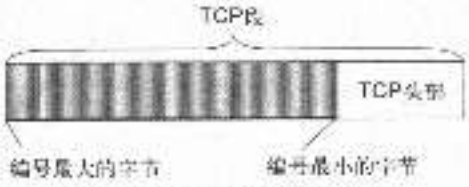


图19-8　TCP段范围内的字节编号

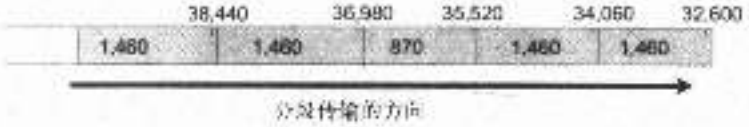


图19-9　序列号和确认号

确认号通常被解释为下一个期待数据字节的编号。在TCP中，只有在数据正确接收之后才发送确认，并不会发送负面的确认。因此，缺乏确认即意味着段被丢失，接收端收到受损段或者确认丢失。

在此协议中，同一个段可能会包含应用程序发送给另一方的数据以及TCP实体确认数据接收的确认消息。

19.2.6　接收端窗口

TCP是一个全双工协议，这意味着双向数据交换的过程被约束在单个连接框架中。每一方同时扮演着发送端和接收端的角色。每方都有一对缓冲区：一个缓冲区用于存储接收段，另外一个用于等待发送的段。除此之外，还有一个缓冲区存储已发送但尚未收到确认的段（图19-10）。

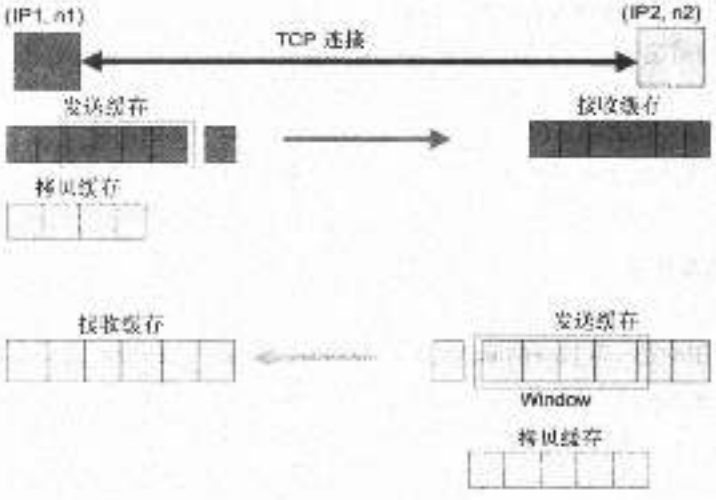


图19-10　TCP连接缓冲区的系统

在建立连接的过程中，以及后来双向传输的过程中，双方作为接收端，相互发送所谓的**接收端窗口**（receiver window）。每一方接收到接收端窗口之后，就知道从接收到上次确认时刻起允许它发送的字节数。换句话说，通过发送接收端窗口，任意一方试着规范自己方向的数据流，通

知另一方它准备接收的字节数（起始于已发送确认的字节号）。

图19-11显示了上一层协议传送给TCP输出缓冲区的字节流。TCP实体切断这个字节流段的顺序并准备将它们送往另一个套接字。在此图中，数据传输方向为从右到左。在此数据流中，可以申明多个逻辑边界。第一个边界将已经发送且确认已经抵达的段隔开，边界的另外一端有一个大小为W字节的窗口。组成该窗口的字节是段已经发送但确认还没有抵达的字节。窗口的余下部分由尚未发送但是因为满足窗口限制能够发送的段组成。最后一个边界申明直到下一个确认到达并且窗口右移后才可以发送的段的起始序列。



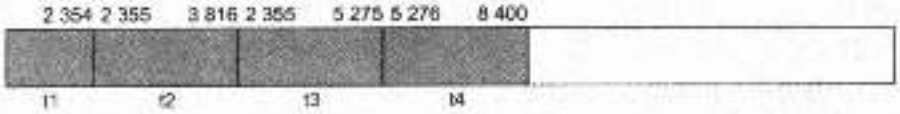
图19-11 TCP中滑动窗口算法的实现

如果窗口大小等于W并且最后收到的确认号的值为N，那么发送端可以一直发送新段直到N + W的字节号落入下一个段。此段超出了窗口的限制；因此，有必要延迟传送直到下一个确认的到达。

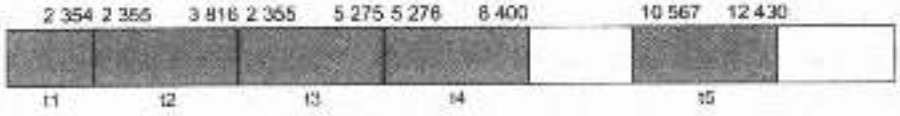
19.2.7 累积确认原则

接收端可以发送确认以同时确认收到多个段，条件是这些段组成了连续的字节流。举个例子（图19-12a），假设缓冲区是密集的、无间隙，且充着多至2 354个字节的字节流。段（2 355－3 816）、（3 817－5 275）以及（5 276－8 400）（括号中的数字代表着每个分段的第一个和最后一个字节），顺序到达缓冲区。这种情况下，接收端为三个分段只发送一个确认就足够了，指明8 401作为确认号。因此，确认过程是累积的。

现在来考虑下一个例子（图19-12b）。段可能会以与发送时不一样的顺序抵达接收端。这意味着接收缓冲区中会存在空隙。打个比方，假设在前文提及的三个段之后，段（10 567－12 430）先于段（8 401－10 566）抵达，而它本应在后者之后到达。发送12 431的确认号并不正确，因为这意味着到字节12 430以前的所有字节都已经收到。既然在字节流中出现了空隙，接收端会重复确认号8 401，以显示它仍然等待着起始于字节8 401的字节流的抵达。从此例中得出结论，与很多其他协议相比，TCP确认连续字节顺序而不是独立数据块的接收。



a）密集缓存填充，在时间点t4传送确认8 401



b）缓存有间隙，在时间点t5重传确认8 401

图19-12 累计确认原则

19.2.8　确认超时

为了安全起见，当TCP发送一个分段到网络时，它将该段的一份拷贝放入重发队列并启动定时器。在该段的确认到达后，拷贝从队列中移除。如果在定时器到期之前确认仍未抵达，就重新发送该段。可能会发生初始段成功传送而重发段也抵达的情况，此时，重复段被丢弃。

确认超时的选择是一个很关键的问题，对它的处理影响着TCP的性能。超时不可以太短；它必须排除无论何时会发生的降低有效系统性能的冗余重传尝试。可是，它也不能太长；它必须避免与不存在或丢失确认相关的长时间空闲时间。

在选择超时值时，有必要考虑物理通信链路的速度和可靠性、它们的长度以及很多其他因素。在TCP中，超时值利用一个复杂的适配算法来定义，主要思想如下：在每个传输过程中，衡量从段发送时刻起到接收到确认期间所流逝的时间。此时间被称为**往返传输时间**（RTT）（round trip time）。RTT的度量值利用随着度量的顺序号增长的权重相互系数的平均而得到，这通过增长最近度量值的影响力来完成。超时值是RTT值乘以某些系数的平均值。实际证明，该系数的值必须超过2，在RTT变化范围大的网络里，在选择超时值时也会考虑到RTT的分散性。

19.2.9　控制接收端窗口

接收端窗口的大小与接收端可用的数据缓存空间有关。因此，接收端窗口通常在不同的连接方有着不同的大小。举个例子，逻辑上我们期待缓存大一些的服务器发送给客户端工作站的将会比客户端发给服务器的接收端窗口更大。根据网络状态，双方可以阶段性地宣布接收端窗口的最新值，将它们动态地递增或递减。

窗口大小的变化影响着网络负载。窗口越大，被发送到网络上的未确认数据部分就可以越多。但是，如果抵达的数据量大于TCP实体所能接收的量，那么该数据将被丢弃。这将导致重传信息的过多尝试以及整个网络特别是TCP软件上无意义负荷的增加。

另一方面，指定窗口太小可能会将数据传输速率限制为每个分段在网络中传输所需的时间。为了避免使用小窗口，有些TCP实现提议数据接收端延缓交换窗口大小直到该连接的可用缓存空间为最大内存空间的20%到40%。然而，发送端必须等到接收端的接收窗口变得足够大才可以发送数据。考虑到这些因素，TCP的开发者提出一种解决办法，在建立连接时定义大窗口而之后大幅度降低其尺寸。还有其他不同的的窗口建立算法，建立连接时选择最小窗口然后当网络成功处理负载之后大幅度提升它的尺寸。

窗口尺寸不仅仅可以被窗口框架内接收数据的一方控制，也可以被发送端调整。如果发送端注册了通信链路的不可靠操作（如，经常延迟传送确认或频繁需要重传时），它可以通过自己的主动权来减少窗口大小。下面规则适用于这种情况：实际窗口尺寸被定为偏小值——由接收端建议的以及发送端提议的值。

中继节点（路由器）和终端节点（计算机）的生成队列说明TCP连接超负载了。在终端节点上缓存溢出时，TCP当发送一个确认时，附带一个新的、尺寸缩小了的窗口。当它同时拒绝接收时，在确认中会申明一个零大小的窗口。然而，即使在这之后，应用程序仍然可以发送消息给拒绝接收数据的端口。为了实现这一点，消息必须被打上紧急标签。在这种情况下，端口被迫接收分段，即使它不得不删除缓存中已存的数据。在收到带有零大小窗口的确认后，发送端协议有时会尝试继续进行数据交换。如果接收端协议准备好了接收消息，它会发送一条带非零大小窗口的确认来响应该查询。

尽管这样描述TCP和UDP很难理解，但它允许您得出结论，也就是说TCP负责着在本来不可靠的网络上确保数据传输可靠性的复杂且重要的任务。

另外一方面，UDP功能的简单性是为了使得其算法简单、规模小、以及操作快速。因此，那

些实现了自己的、足够可靠的面向连接消息机制的应用程序宁愿使用可靠性稍低但是传输更快的工具来在网间传输数据。与TCP相比，UDP正是这样的工具。如果高速通信链路确保了足够的可靠性等级，而无需建立逻辑连接，无需确认传输的分组，也可以用UDP。最后，因为TCP是面向连接的，与UDP比起来，它不能用于广播或多播的数据传输。

19.3 路由协议

19.3.1 路由协议的分类

自动创建路由表保证了网络上传输分组路由的高效性；因此，可以采用不同标准来选择路由。当前IP网络使用选择最短路由的路由协议。这种情况下，分组传输的距离被解释为中间节点（路由器）的个数，通常称为跳数。也可以综合考虑到连接路由器链路额定带宽、链路可靠性、链路延迟等因素。

一个路由协议必须在路由表中创建一致的路由表。路由表确保了从源网络到目的网络间分组在有限步数内完成传送。也可以想像一下不一致的两张表，此时，路由器1的路由表申明送往网络A的分组必须被送往路由器2，而路由器2的路由表可以发送同一分组给路由器1。当前的路由协议确保了表之间的一致性；然而，这个属性并不是绝对的。举个例子，如果网络中发生了改变，比如通信链路或路由器崩溃，可能会出现由不同路由器间的路由表缺乏协调而引起的网络操作不稳定的时期。通常，路由协议需要一点时间，称为**聚合时间**（convergence time），期间所有的网络路由器在多次迭代交换辅助信息之后，将必要的变化加入自己的路由表中。此操作的结果是，所有路由表重新处于一致的状态。不同的路由协议具有不同聚合时间值。

对于因特网上的成功操作，网络技术必须是**可延拓的**（scalable）。这意味着它们必须保证某种形式的层次化运用。因特网路由遵从这个原则并将因特网划分为**自治系统**（autonomous system）。因此，因特网中的路由具有明确的层次特性。

任何现存的路由协议都可以在自治系统中使用，但是，在自治系统之间必须使用相同的协议，作为一种通用的语言，如Esperanto即在自治系统中用来彼此之间相互通信。

在IP网络中，**内部网关协议**（interior gateway protocol，IGP）在自治系统中扮演的角色授权给了下述三个协议——**路由信息协议**（routing information protocol，RIP）、**开放最短路径优先**（open shortest path first，OSPF）以及**中间系统到中间系统**（intermediate system to intermediate system，IS-IS）。另外一种协议类型，叫做**外部网关协议**（exterior gateway protocol，EGP），包含用于在自治系统间选择路由的路由协议。如今，此任务由**边界网关协议**（border gateway protocol，BGP）来完成。

1. 不带表的路由（tohere）

在我们继续探讨路由协议分类之前，有必要指出因特网中也存在着不需要路由器中存在路由表的分组传递方法。

洪泛路由（flood routing）是网络中传递分组的最简单的方法。此时，每个路由器将分组传递给除了刚递交分组给它的节点之外的最近邻居。很显然，这个方法效率最低，网络带宽都被浪费了。但是，这个方法是可行的，因为网桥和交换机对付未知地址的帧正是采用这种方法传递的。

另外一个不带路由表路由的例子是**基于事件的路由**（event-dependent routing），根据之前成功传送的经验，分组沿着同样的路由传送给特定的目的网络（对于给定目的地址）。该路由方法在因特网形成时使用。根据该方法，在分组发送之前，ICMP的echo请求被送往所有的或多个邻居；然后，基于收到的echo回复时间，选择拥有最小响应时间的那个邻居。

这个方法适用于面向连接协议的网络。建立连接的请求可以同时送往多个邻居，而确认必须

被送往第一个发送回复的邻居。

还有一个不带路由表路由的类型是**源路由**（source routing）。这种情况下，发送端在分组信息中放入分组传送到目的网络过程中所必须经过的中继路由器信息。基于该信息，每个路由器获取下一个路由器地址，并且如果它是最近邻居的地址，就将分组传递给它以进行下一步处理。决定分组沿网络传送的确切路由问题至今尚未解决，路径可以由管理员人工指定也可以由发送节点自动决定，后者必须支持某一路由协议。这里的路由协议必须通告发送端有关网络拓扑和状态信息。源路由在早期因特网实验中测试过，一直被保留为IPv4的未启用选项。在IPv6中，源路由是分组传递的标准模式之一，并且存在着实现此模式的特定头部定义。

2. 自适应路由

当在表基础上执行路由时，有两种不同的方法：*静态路由和自适应*（动态）路由。

在**静态路由**（static routing）时，路由表由*网络管理员手动*创建并加入到每个路由表中。路由表中的所有记录有着静态状态，意味着它们保持永远有效。当某些网络元素状态发生改变，管理员必须手动输入对路由表进行适当的修改。比如说，可能需要更改分组的路由。这必须尽快完成；否则，网络上就会出现错误。

自适应路由（adaptive routing）在网络配置改变时*自动更新路由表*。更新路由表正是路由协议的任务。这些协议基于各个算法操作，这些算法允许所有路由器在网络中的链路拓扑上收集信息并且灵活反应链路配置的所有变化。如果使用自适应路由，路由表通常包含一定期间内每个独立路由都保持有效的信息。此时间段被称为路由的*生存时间*（*TTL*）。当TTL期限过期并且路由协议尚未确认路由的存在时，该路由被认为是不起作用的，分组不会沿此路由传送。

路由协议的分类有分布式和集中式两种。

- 当使用**分布式方案**（distributed approach）时，网络不会包含任何专门收集和总结网络拓扑信息的路由器。相反，此项工作分布在网络的所有路由器中。每个路由器在从其他网络路由器的路由协议收到的数据基础上来构建自己的路由表。
- 当使用**集中式方案**（centralized approach）时，网络中有一个专属路由器。它从其他路由器收集网络拓扑和状态的信息。然后该专用路由器，有时叫做路由服务器，选择一个可行的行为方案。它可以为所有剩余的网络路由器构建路由表，然后发布到网上，这样每个路由器可以获取自己的拷贝或路由表；随后，可以根据自己的路由表来决定分组传送的路径。

IP网络当前使用的路由协议被分类为分布式自适应协议。

自适应路由算法必须满足几个重要的需求。首先，这些算法选择的路由即使不是最优也必须是高效的。其次，算法必须足够简单，因为它们的实现不能浪费网络资源。特别是，它们不能需要大量计算或产生过多的控制流量。最后，路由算法必须具备收敛性——即它们必须总是协调路由表的构建以保证在合理时间内覆盖到所有的网络路由器。

当前计算机网络中使用的自适应路由算法被分成两组，每个实现下列类型算法之一：

- 距离向量算法
- 链路状态算法

3. 距离向量算法

在**距离向量算法**（distance vector algorithm，DVA）中，每个路由器阶段性地广播向量，其中含有该路由器到它所知各个网络的距离。路由协议发送的分组通常称为**广告**（advertisement），因为路由器利用它们来通知网络结构上所有知道的其他路由器。通常，DVA中的距离被解释为跳数。然而，另外一种度量方式也是可行的，它不仅考虑到中继节点（路由器）的个数同时也考虑连接两个相邻路由器的链路带宽。

从邻居接收到向量之后，路由器将向量中指定的距离加上自己到该邻居的距离并且补充上自己所知道的其他网络信息，添加到该向量。当它们连接在当前路由器端口上时，这些网络信息可能是直接获取的、或者来自于其他路由器发送的类似广告。然后，路由器将新的向量值广播到网络上。最后，每个路由器将从它的邻居信息中得知连接到因特网上的所有网络以及到它们的距离。至此，它将从到每个网络的多个可选路由中选择度量值最小的一个。传递过关于此路由信息的路由器在路由表中被标为**下一跳**（next hop）。

DVA只在小型网络中效率比较高。在大型网络中，它们的大量阶段性通信量给通信链路带来沉重的负担。而且，根据该算法不能很好地处理配置变化，因为路由器并没有网络链路拓扑的准确信息。相反，它们只有大概的信息，距离向量；更何况，这不是一个直接接收的中间信息。

在基于DVA的协议中，RIP是使用最广泛的。该协议有两个版本——RIP IP，为IP服务；以及RIP IPX，与IPX一起工作的。

4. 链路状态算法

链路状态算法（link state algorithm，LSA）为每个路由器提供足够构建网络链路的准确图的信息。所有的路由器在同一个图的基础上操作，使得路由处理面对配置变化时更加稳定。

每个路由器利用网络图找到抵达因特网中每个子网的路由，根据某些特定标准这些路由都是最优的。

为了发现连接到端口上的通信链路的状态，路由器阶段性地与最近邻居交换短小的*HELLO*分组。

链路状态上的广播并不是像DVA协议那样阶段性地重复。与之相反，它们仅仅在交换*HELLO*消息时发现某一链路状态改变时才传送。因此，LSA协议产生的控制流量远比DVA协议产生的要少得多。

LSA协议的实例有OSI栈的IS-IS协议（此协议也用于TCP/IP栈中）、TCP/IP栈的OSPF协议以及Novell栈的NetWare链路服务协议。

5. 使用多个路由协议

多个路由协议可以同时在一个网络中操作（图19-13），这意味着在一些（但是不需要是全部）网络路由器上，安装并运行着多个路由协议。很显然，只有类似名字的协议才利用网络进行交互。这意味着如果路由器1支持RIP和OSPF协议，路由器2只支持RIP，路由器3只支持OSPF，那么路由器1将通过RIP与路由器2交互，通过OSPF与路由器3交互；而路由器2和3将不能够直接通信。

在支持多个协议的路由器中，路由表中的每条记录都是这些协议之一的操作结果。如果特定网络上的信息由多个协议提供，那么为了保证路由选择的确定性，会设置路由优先级；否则，不同协议的数据可能会导致不同的有效路由。通常，会倾向于LSA协议因为与DVA协议比起来，它们拥有更全面的信息。在有些操作系统里，显示或打印路由表的表格包含指定用于获取每个记录的路由协议的特殊掩码。即使该掩码没有显示，它也总是存在于路由表的内部标识中。默认情况下，运行于特定路由器上的每个路由协议只发布根据该协议路由器获得的信息。因此，如果路由器从RIP得到一些网络的信息，它将使用同样的协议在网络上发布该路由的广告信息。

然而，有人可能会问这样的问题："路由器怎样支持不同的路由协议，使得它们在因特网中交换路由信息让所有组成网络都可以访问得到？"为了使路由器能用利用一种路由协议来发布利用另外一种路由协议接收到的路由信息，需要建立特定的内部操作模式，通常叫做*再分配模式*（*redistribute mode*）。这种模式确保特定协议不仅可以利用它路由表中的"本地"记录，也可以利用通过配置过程中指定的其他路由协议获得的记录。

正如描述中可以看出，在一个网络中使用多个路由协议不是一件简单的事。管理员必须执行配置每个路由表的特殊任务。对于大型异构网络来说，需要原则性的不同解决方案。

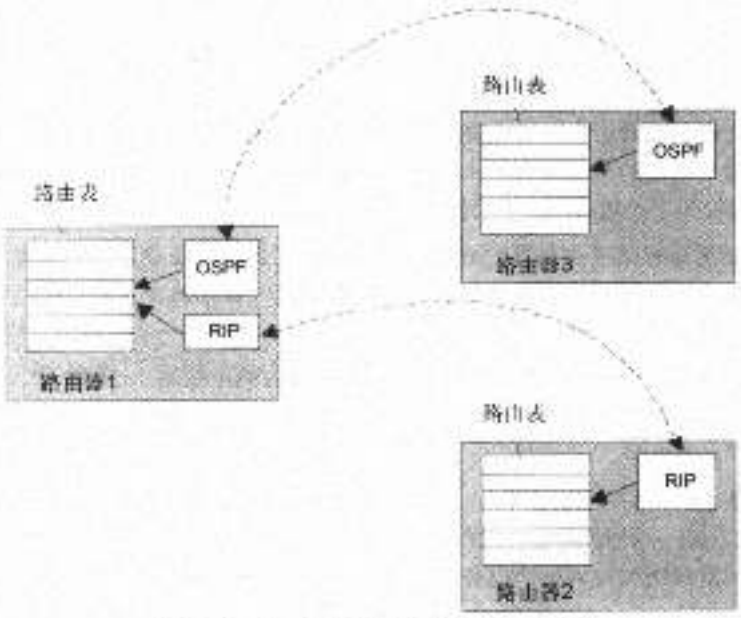


图19-13　同一网络中多个路由协议的操作

这种解决方案曾经应用在因特网上，当今最大规模的异构网络。

6. 外部与内部网关协议

除了第5章中描述的因特网组织结构以及将因特网划分为属于不同ISP的子网以外，因特网还可由自治系统组成。

一个自治系统（*autonomous system*）是由公共监督管理连接成的一组子网，其确保所有路由器的公共路由策略都被包含在自治系统之中。通常，自治系统由单个ISP来控制。该ISP决定在特定自治系统中使用哪些路由协议以及路由信息在这些系统如何分布和重新分布。顶级ISP和大型公司可以将他们的网络表示为多个自治系统的集合，自治系统注册采用与IP地址和DNS域名的注册方法相类似的集中方式。

所有的自治系统都是统一编号的，自治系统的号码包含16位，它与组成自治系统网络的IP前缀无关。

根据其概念，因特网看似一个连通的自治系统集合，其中每个都包含着相关的网络（图19-14）。

将因特网划分为自治系统的主要目标是确保路由的多级解决方案。在引入自治系统之前，路由采用两级解决方案——网络层路由在节点组（网络）之间进行，网间路由则由低层技术来实现。这意味着路由定义了在网络中传递的顺序。

随着自治系统的到来，出现了第三级路由——首先在自治系统层选择路由，然后在它们的组成网络层。

自治系统由外部网关[⊖]（exterior gateway）相连。重要的是在外部网关之间只允许存在一种路由协议。而且，不能是任意一种协议。相反，这必须是因特网委员会为外部网关选用的标准协议。这种路由协议被叫做外部网关协议（exterior gateway protocol，EGP）。现今情况下，只存在着一种这样的路由协议，BGPv4。所有其他的协议（如，RIP、OSPF、IS-IS）都属于内部网关协议（interior gateway protocol，IGP）。

⊖　从现在开始，词汇路由器和网关被视为同义词，继承传统因特网术语的同时也不忘使用更新的术语。

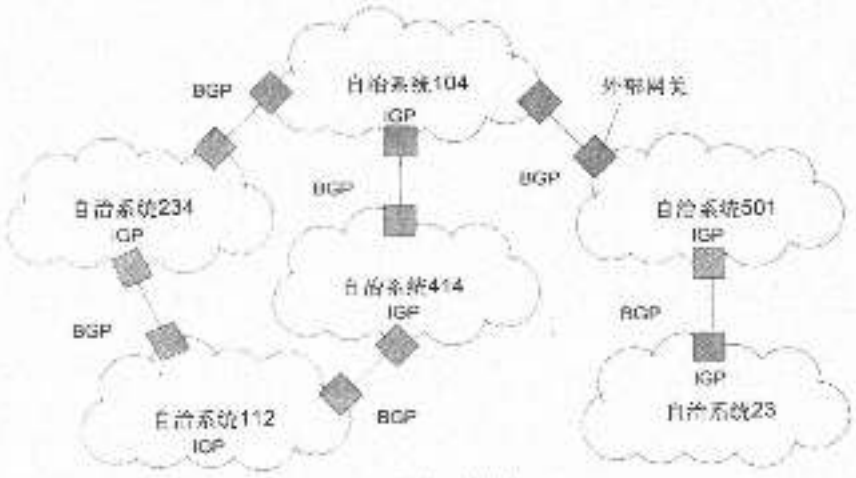


图19-14 因特网的自治系统

EGP负责选择路由作为自治系统序列（*sequence of autonomous system*）。作为下一个路由器结果，指定了下一个自治系统的访问点地址。

IGP负责自治系统内部（*within an autonomous system*）的路由。自治系统指定了准确的路由器系列，从访问点到路由离开该自治系统点。

自治系统形成了因特网主干。自治系统的概念使得因特网主干管理员不用关心在低级网络层上出现的分组路由问题。对于主干网管理员来说，自治系统内部使用的路由协议根本无关紧要，唯一有关系的路由协议就是BGPv4。

19.3.2 路由信息协议

1. 建立一张路由表

路由信息协议（routing information protocol，RIP）是一个基于DVA算法的IGP。它是最早的路由协议之一，得益于其实现的简单性，一直在计算机网络中广泛使用。

对于IP网络，存在着两个版本的RIP：RIPv1和RIPv2。第一个版本，RIPv1，不支持掩码。与RIPv1相比，RIPv2使用网络掩码，因此，它更适应如今的需求。然而，虽然构建RIPv2使用的路由表的过程与RIPv1并没有本质上的区别。为了简单起见，我们只描述第一个版本的过程。

RIP版本允许使用不同类型的度量方法来定义到网络的距离。比如说，可以使用最简单的度量方法，跳数，或者复杂一点考虑到网络带宽、延迟和网络可靠性（即，对应于IP分组的*ToS*字段中的*D*、*T*和*R*标志）的度量方法，以及这些度量方法的组合。度量方法必须具备加和性的特征，也就是，组合路由的度量必须等于该路由组成特征的各种度量之和。在RIP的大多实现中，使用的是最简单的度量，即跳数（也就是，分组传递到目的网络必须经过的中转路由器）。

现在让我们来看看在图19-15所示的网络示例中利用RIP构建路由表的过程。

步骤1. 创建最小表

考察中的网络包含了四个路由器连接的八个

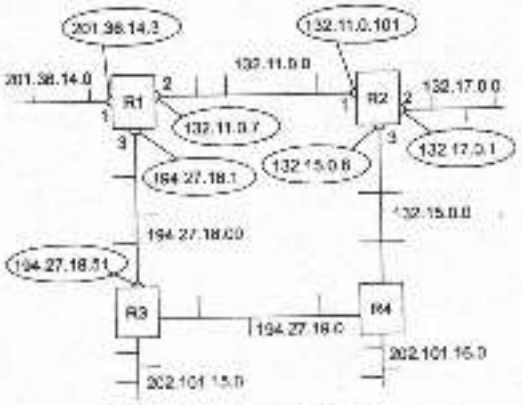


图19-15 建在RIP网关上的网络

IP网，这四个路由器的标识符为：R1、R2、R3和R4。根据RIP操作的路由器必须有标识符。然而，对于协议的操作，标识符并不是必需的，因为它们不在RIP消息中传递。

初始状态下，TCP/IP栈软件自动创建**最小路由表**（minimal routing table），它只考虑直接相连的网络。在图19-15中，路由器端口地址被放置在椭圆中以示与网络地址的区别。

从表19-1中，可以预览一下R1路由器的最小路由表。

其他路由器的最小路由表看起来与之类似，举个例子，R2路由器的路由表将包含三条记录（表19-2）。

表19-1　R1路由器的最小路由表

网络号	下一个路由器的地址	端口	距离
201.36.14.0	206.36.14.3	1	1
132.11.0.0	132.11.0.7	2	1
194.27.18.0	194.27.18.1	3	1

表19-2　R2路由器的最小路由表

网络号	下一个路由器的地址	端口	距离
132.11.0.0	132.11.0.101	1	1
132.17.0.0	132.17.0.1	2	1
132.15.0.0	132.15.0.6	3	1

步骤2. 向邻居发送最小表

初始化之后，每个路由器开始向邻居发送包含最小表的RIP消息。

RIP消息在UDP分组中传送并且包含每个网络的两个参数：其IP地址以及从传递此消息的路由器到网络的距离。

邻居是当前路由器可以直接传递IP分组而不需用使用中转站的路由器。比如说，对于R1路由器，邻居是R2和R3，而对于R4，则为R2和R3。

因此，R1路由器传送下面的消息给R2和R3：

网络 201.36.14.0，距离1

网络 132.11.0.0，距离1

网络 194.27.18.0，距离1

步骤3. 从邻居接收RIP消息并且处理路由信息

在从路由器R2和R3收到类似消息之后，R1路由器将每个收到的度量递增1并通过收到消息的那个路由器的端口“记录”下来。如果这条记录被记入路由表中，该路由器地址将变成下一个路由器地址。然后，路由器开始对比新消息与存储在路由表中的记录（表19-3）。

表19-3　往R1路由器的路由表中增加记录

网络号	下一个路由器的地址	端口	距离
201.36.14.0	201.36.14.3	1	1
132.11.0.0	132.11.0.7	2	1
194.27.18.0	194.27.18.1	3	1
132.17.0.0	132.11.0.101	2	2
132.15.0.0	132.11.0.101	2	2
194.27.19.0	194.27.18.51	3	2
202.101.15.0	194.27.18.51	3	2
~~132.11.0.0~~	~~132.11.0.101~~	~~2~~	~~2~~
~~194.27.18.0~~	~~194.27.18.51~~	~~3~~	~~2~~

从邻居路由器收到编号为4到9的记录，它们是插入表中的候选记录。然而，只有记录4到7被输入到路由表中；记录8和9并没有。这是因为这些记录包含R1路由器的路由表中已有的数据，而且申明的距离比已有记录中的距离还要长。

RIP只在新消息的度量优于现存值时才会更换掉特定网络的记录。也就是说，路由表只为每个网络保存一条记录。即使存在着到某一网络距离相等的多条路径，也只有一条记录保存在路由表中，即第一个抵达此路由器的那个。此规则有一个例外：如果收到来自同一路由器关于某一特定网络的更差消息，而表中当前记录是基于此路由器而建的，那么这条更差度量的记录将取代表中的当前记录。这是因为网络环境变差了，而路由器诚实地汇报了这一点。

其他网络路由器对于收到的新消息执行类似的操作。

步骤4. 向邻居发送一个新表

每个路由器向邻居发送它最新的RIP消息。在此消息中，放置着它所知道的所有网络数据，包括与之直接连接的网络以及从其他路由器的RIP消息中获取的远程网络。

步骤5. 从邻居接收新的RIP消息并且处理收到的信息

步骤5重复了步骤3的工作——所有路由器接收RIP消息并且进行处理。然后，在该信息的基础上，它们更新自己的路由表。

来看一下R1路由器是如何执行这个操作的（表19-4）。

表19-4 更新R1路由器的路由表

网络号	下一个路由器的地址	端口	距离
201.36.14.0	201.36.14.3	1	1
132.11.0.0	132.11.0.7	2	1
194.27.18.0	194.27.18.1	3	1
132.17.0.0	132.11.0.101	2	2
132.15.0.0	132.11.0.101	2	2
~~132.15.0.0~~	~~194.27.18.51~~	~~3~~	~~3~~
194.27.19.0	194.27.18.51	3	2
~~194.27.19.0~~	~~132.11.0.101~~	~~2~~	~~3~~
~~202.101.15.0~~	~~194.27.18.51~~	~~3~~	~~2~~
202.101.16.0	132.11.0.101	2	3
~~202.101.16.0~~	~~194.27.18.51~~	~~3~~	~~3~~

在该步骤，路由器R1从路由器R3接收到关于132.15.0.0网络的信息，该信息为前一步骤中从路由器R4所接收。路由器R1已经获知132.15.0.0网络，而且早期的信息中度量更低；因此，此网络的新信息被丢弃。

有关202.101.16.0网络的信息被路由器R1首次接收。来自两个邻居R3和R4的该网络相关的信息同时到达。既然两条消息中度量相等，率先抵达的记录将被记入路由表中。在此例中，路由器R2在R3前面，是第一个发送RIP消息给R1的路由器。

如果路由器阶段性地重复发送和处理RIP消息的步骤，那么在有限期间内，就可以在网络中建立正确的路由模式。这里，在正确的路由模式下，将会建立路由表使得利用某有效路由可以从任意网路抵达其他所有的网络。分组将抵达目的地并且不会丢失在类似于R1-R2-R3-R4的循环路由里（图19-15）。

如果所有的网络路由器、它们所有的接口以及连接它们的所有通信链路总是可操作的，那么可能很少根据RIP发送广告，比如，每天一次。但是，网络结构的变化总是在发生着，路由器和通信链路的可用性也会改变，现存网络中会增加或移除新的路由器和通信链路。

为了适应网络结构中的这些变化，RIP使用了好多种机制。

2. 使RIP路由器适应网络状态的变化

RIP路由器很容易适应新的路由，它们只是将下一个消息中的新信息传递给邻居。因此，新信息逐渐地被网络中所有路由器所获知。然而，适应与路由丢失相关的负面结果更加困难。这是因为RIP消息不包含指定到某网络的路径不复存在的字段。

下面两个通知机制用于通知路由器某些路由不再有效的信息：

- 路由TTL过期
- 通知到网络的某一特定距离不再有效

为了实现**路由TTL过期**（expiration of the route TTL）的机制，路由表根据RIP收到的每条记录都有一个有限的TTL。当另一个RIP消息抵达时，确认特定记录的有效性，TTL计时器被重置到初始值然后每秒钟递减一。如果这个路由上的新消息在超时期限内没有到达，那么该路由即被标为无效。

超时期限与网络上广播向量的间隔相关。在RIP中，该间隔为30秒，超时值为此间隔长度的6倍（即180秒）。需要保留六倍时间以确认特定网络已经变得无效并且不是因为RIP消息丢失而造成的连接中断。注意RIP消息的丢失是可能发生的，因为RIP使用UDP传输协议，而后者并不保证消

息的可靠传输。

如果任何路由器发生故障并停止发送有关通过它可抵达网络的信息，那么在180秒之后，该路由器产生的所有记录将被它最近的邻居宣布无效。之后，此过程将在最近邻居的邻居上重复。此时，经过360秒之后会丢弃掉无效记录，因为在最初180秒期间故障路由器的最近邻居仍然传输着它的有关信息。

关于因为路由器故障而失效的网络信息在网络上缓慢地扩散着。因此，广播间隔被定为30秒的值。

超时机制（timeout mechanism）工作于路由器不能通知邻居有关失效路由的信息时，可能是因为它本身也发生了故障，还可能是因为传输消息的通信链路发生故障。

在可能发送消息时，RIP路由器并不会在消息中使用任何特殊属性。取而代之的是，它们指定了到该网络的无限距离。注意在RIP中，该距离为16跳。如果使用了其他度量而不是跳数，需要指定该度量的一个无限值。考虑一下当路由器收到一条某一网络距离设为无限值（16跳）的消息时会发生什么。注意如果该网络距离为15跳，结果也一样，因为路由器会将收到的值加1。收到这样的消息，路由器必须检查这个“负面”消息是否来自于曾经在路由表建立该记录时发送消息的那个路由器。如果的确如此，那么此消息被认为是可靠的，该路由被标为无效。

为“无限”距离选择如此小的一个值，是因为某些情况下链路故障引起RIP路由器错误操作的期限延长。错误行为表现为网络环中的分组无限循环。无限值设得越小，这种错误网络操作的时间就越短。

示例　考虑一下图19-15中所示网络上的分组循环。

假定路由器R1发现它到直接相连网络201.36.14.0的连接丢失。比方说，这可能是因为接口201.36.14.3故障而引起的。路由器R1在路由表中将网络201.36.14.0标为无效的。更坏的情况是，路由器在发送了常规预定的RIP消息之后立刻发现了该事实。这时，在广播新一轮循环开始之前有将近30秒时间供它通知邻居到网络201.36.14.0的距离已经变为16。

每个路由器在内部计时器的基础上操作，而不会同步它发送给其他路由器的广播。因此，很有可能路由器R2会抢在R1之前，在R1有时间发送网络201.36.14.0已失效的消息之前传送出它的消息。路由器R2发送的消息可能包含根据其路由表中记录产生的数据（表19-5）。

该记录来自于路由器R1并且直到接口201.36.14.3失效之前一直都是正确的。现在这条记录不再有效；可是，路由器R2还没有被通知到。

现在，路由器R1将收到关于201.36.14.0网络的新信息，根据该信息，此网络通过路由器R2距离2范围内可达。在此之前，R1也从路由器R2收到过同样的信息。然而，路由器忽略了此条信息因为它自己的度量距离更短。现在，R1必须收到来自于R2有关201.36.14.0网络的数据，取代路由表中的记录，将此网络标为无效的（表19-6）。

表19-5　路由器R2的路由表中的记录

网络号	下一个路由器的地址	端口	距离
201.36.14.0	132.11.0.7	1	2

表19-6　路由器R1路由表中的记录

网络号	下一个路由器的地址	端口	距离
201.36.14.0	132.11.0.101	2	3

这样，网络中产生了一个路由循环：发送到网络201.36.14.0的分组将被路由器R2传递到路由器R1，并且路由器R1会将它们返回到路由器R2。IP分组在此循环中轮转直到每个分组的TTL过期。

路由循环将在网络存在相对比较长的时间，现在来考查路由表记录TTL的多倍的时间期限：

- 0到180秒。在接口失效之后，错误的记录仍然存在于路由器R1和R2的路由表中。路由器R2继续向路由器R1提供它关于网络201.36.14.0距离为2的记录，因为该记录的TTL还没有过期。

因此，分组将陷入循环中。

- 180到360秒。此阶段初始时，关于网络201.36.14.0距离为2的记录的TTL将在路由器R2上到期。这会发生是因为R1在前一阶段发送网络201.36.14.0距离更大的广告，而且它们无法确认此条记录。现在，R2从路由器R1接收了一条网络201.36.14.0距离为3的记录，它将此记录中的距离转化为4。另一方面，R1并没有收到来自于R2有关网络201.36.14.0距离为2的新消息；因此，此记录的TTL开始递减。分组继续在循环中轮转。
- 360到540秒。路由器R1上有关网络201.36.14.0距离为3的记录TTL过期。路由器R1和R2在此互换角色——现在R2为R1提供有关到网络201.36.14.0路径的过期信息；然而，这次，距离变成了4，路由器R1递增1之后得到5。循环中分组轮转继续进行。

如果不是“无限”距离选择为16，这个过程会无限制的继续下去。更准确地说，它将继续直到超出距离字段的长度，这样在下次尝试增加距离时便会发生溢出。

最后，前面描述过程的下一步中路由器R2将从路由器R1返回距离值15，递增之后，即变成了16。路由器R2然后会注册该网络为不可及。网络不稳定操作期间持续36分钟！

15跳的限制局限了RIP对网络的应用领域，只适用于中转路由器数不超过15的网络。而对于稍大一些的网络，有必要使用其他路由协议，比如OSPF，或者将网络分割为自治区域。

前面描述的例子说明了根据RIP工作的路由操作不稳定的主要原因。原因在于DVA协议的基本原则，利用从第三方收到的信息。路由器R2在无法确认可靠性的情况下将网络201.36.14.0可用的信息传递给R1。

说明 路由循环并不是在接口或路由器失效时产生。如果路由器R1有时间在从R2收到过时消息之前将网络201.36.14.0不可用的信息传递出去，那么就不会产生路由循环。有必要提到，平均说来，即使在采取任何特殊措施防止它们之前，路由循环发生的概率也不会超过50%。减少路由循环的方法将在下一节中描述。

3. 在RIP中消除无效路由的方法

尽管当一些路由器使用不存在路由的过时信息时，RIP无法完全消除网络中的瞬间状态，但仍存在着大多数情况下可以解决这种问题的方法。

前面小节中提到路由循环在两个邻居路由器之间产生的情况可以利用**水平分割**（split horizon）的方法来妥善解决。该方法的前提是暗示存储在特定路由器路由表中的有关某一网络的路由信息从不发送给接收此信息的路由器（就是当前路由的下一个路由器）。

事实上，当前所有根据RIP协议操作的路由器，都使用水平分割技术。如果前面示例中的路由器R2支持水平分割技术，它就不会将网络201.36.14.0的过时信息发送给R1路由器，因为它曾经从R1路由器上收到此信息。

然而，当路由循环由超过两个路由器引起时水平分割将不起作用。再仔细考虑一下图19-15中所示网络的情况，当路由器R1丢失其到网络201.36.14.0的连接时。假设此网络的所有路由器都支持水平分割技术。此时的R2和R3不会返回网络201.36.14.0距离为2的数据给R1，因为这是给它们发送此信息的路由器。但是，它们会彼此之间继续传递网络201.36.14.0距离为4有效的信息，因为它们从组合路由收到的该信息，而不是直接从路由器R1。比如说，路由器R2从R4-R3-R1链收到此信息。因此，路由器R1可能再次被欺骗直到R3-R4-R2链中的每个路由器都丢弃网络201.36.14.0可用的记录。

为了防止分组在链路失效后在复杂环中循环，存在着两种其他的方法，叫做触发更新（*triggered update*）和压缩（*hold down*）。

利用触发更新技术（*triggered updates technique*），路由器在收到关于到特定网络的度量改变的数据之后，并不等待超时期限过期才更新路由表信息。取而代之的是，它把路由变化的数据立即传送出去。很多情况下，此技术可以防止有关失效路由过时信息的传送。但是，它的控制信息增加了网络的负载；因此，触发更新经过一定的延迟后执行。正因为这点，某一路由器中常规更新比来自前一个路由器的触发更新到达要稍微早一点发生。因此，此路由器仍然来得及向网络中传送关于不存在路由的过时信息。

压缩技术（*hold down technique*）则减少这种情况的发生。该方法引入从近期不可及的网络上接收新信息的超时机制。超时防止路由接收那些距离故障链路有一定距离并且传送过时数据的路由器传来的过时信息。通常认为在压缩阶段，这些路由器会从它们的路由表中删除此路由，因为它们不会再收到相关的新消息了。这样，它们便不会在网上传播过时信息。

19.3.3　开放最短路径优先

开放最短路径优先（OSPF）协议是LSA算法当前的一个实现，它于1991年被采用，特征是拥有很多面向大型、异构网络而设计的能力。

1. 建立路由表的两个步骤

在OSPF中，构建路由表的过程被分为两大步骤。第一步，每个路由器构建网络链路图；图中顶点为路由器和IP网，边为路由器接口。为了实现这一点，所有路由器和邻居交换它们所配置的网络图。这个过程类似于RIP中扩散距离向量的过程，然而，与被扩散的信息有着本质上的区别，这次是网络拓扑相关的信息。路由器交换的消息被称为**路由器链路广告**（router links advertisement）。而且，在传输拓扑信息时，路由器并不会改变它，正如RIP路由器所做的那样。作为散布拓扑信息的结果，所有路由器的网络图都保持一致。此信息被存储在路由器的**拓扑数据库**（topological database）中，又称为**链路状态数据库压缩技术**（link state database hold down technique）。

第二步为利用接收到的图来选择最优路由。每个路由器将自己看作是网络中心并且找寻到达每个可及的网络的最优路由。根据网络图寻找最优路径的问题相当复杂而且耗费资源，为了解决这个问题，OSPF实现了Dijkstra的迭代算法。在使用此算法所找到的每个路由中，只记忆一步，即，根据单跳路由原则算出到下一路由器的跳数。该步的数据被输入到路由表中。如果到目的网络多个路由有着同样的度量，那么路由表将储存这些路由的第一步。

2. HELLO路由广告

在构建了原始路由表之后，有必要跟踪网络链路状态并将更新输入到路由表中。OSPF路由器并不会交换所有信息来控制链路状态，如效率低下的RIP路由器那样。取而代之的是，它们传输特殊的、简短的*HELLO*消息。如果网络状态并没有改变，OSPF路由器不会更新路由表，也不会向邻居发送链路广告。如果连路状态发生了改变，路由器向最近的邻居发送一则新的广告，此广告只与状态改变了的链路相关。很自然，这个解决方法占用的网络带宽更少。收到链路状态改变的新广告之后，路由器重建网络图并且重复搜索最优路径的过程。注意，后一步并不适用于所有的路由：只有被改变所影响到的路由才被重新计算。完成这些之后，路由器更正了它的路由表。同时，路由器重新传输广告给最近的邻居，当然要排除发送给它此广告的那个邻居。

当网络中出现新链路或新邻居时，路由器从新的*HELLO*消息中收到此信息。虽然*HELLO*消息的尺寸相对较小，但它们仍然含有发送此消息的路由器的详细信息以及有关它最近邻居的数据，这使得此路由器被准确无误地识别出来。*HELLO*消息每隔10秒发送一次以增加路由器对网络上发生所有变化的接收速度。这些消息的小尺寸使得频繁测试网络邻居及它们之间的链路成为可能。

3. 度量

通常，OSPF使用考虑了网络带宽的度量。除此以外，还可以使用考虑IP分组中QoS需求的另外两种度量，即分组传输延迟和分组传输可靠性。对于所使用的每种度量，OSPF都会建一个单独的路由表。所需路由表的选取根据抵达路由器分组的QoS需求来执行（图19-16）。

这种网络的结构反映在图19-17的图中。

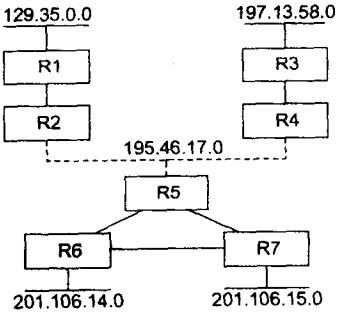


图19-16 网络分片

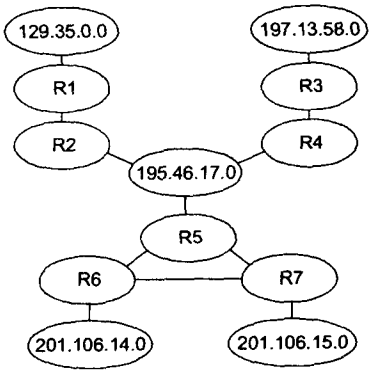


图19-17 利用OSPF协议构建的网络图

路由器连接到LAN并通过“点到点”的WAN链路相互直接相连。在它的广告中，OSPF发布关于下列两种类型链路的信息：路由器到路由器以及路由器到网络。第一种链路的例子有“R3-R4”连接，连接“R4-195.46.17.0/24”是第二种链路的例子。注意，R3和R4也代表着IP地址；然而，我们使用符号标识符来将图中的点与网络区分开来，对于网络我们保留了IP地址的标准表示法。如果点对点连接也被分配了IP地址，它们将变成图中新增的顶点，像LAN那样。网络掩码的信息也随同IP地址一起传送。

与RIP路由器相似，在初始化之后，OSPF路由器只有直接相连网络的链路信息，它们开始把消息扩散给邻居。与这个过程同步进行的是，它们通过接口发送*HELLO*消息这样路由器可以立即获知最近邻居的标识符。该消息补充了它的拓扑数据信息。随后，拓扑信息开始在邻居网络间扩散，经过一段时间后，到达了最远的路由器。

每个链路都有一个度量。OSPF协议支持很多协议使用的标准度量，比如STA。这些值反应了真实的网络性能；对于以太网，该值为10单元；对于快速以太网，为1单元；对于T1通道，为65单元；对于56Kb/s通道，为1 785单元。在使用诸如千兆以太网或STM-16/64的高速链路时，管理员必须为最高速链路指定另外一种适当距离单元的速度规模。

当在图上选择最优路径时，有必要考虑到图上与每条边相关的度量。当路径中包含特定边时，该度量被加入到路径度量中。比如，在前面提到的例子里，路由器R5通过T1链路连接到路由器R6和R7，而路由器R6和R7之间通过56Kb/s链路连接。此时，R7会找到一条到201.106.14.0网络的最优路径作为组合路由，它首先经过路由器R5然后路过R6，因为该路由器的路由度量为65+65=130个传统单元。直接通过R6的路径不是最优的，因为路由度量为1 785。当使用跳数时，将选择直接经过R6的路由，尽管这并不是最优的。

OSPF协议允许在路由表中存储到同一网络的多个路由，假设这些路由具备同样的度量。如果这种记录存在于路由表中，那么路由器将通过切换发送分组的路由来实施负载均衡模式。

4. OSPF稳定性

拓扑数据库中的每条记录都有它自己的TTL，正如RIP的路由记录一样。链路状态的每条记录都有它用于控制记录TTL的计时器。如果从另一个路由器收到的路由器拓扑数据库中的任

何记录变得过时无效时，路由器可以利用OSPF协议的*特殊链路状态请求*消息来申请此记录的另外一份拷贝。对于这条消息，路由器必须从直接测试请求链路的路由器接收到*链路状态更新*的回复。

初始化路由器时，为了更加可靠地同步拓扑数据库，它们阶段性地交换数据库的所有记录。然而，所有数据库记录的交换时间远远长于RIP路由器的类似时间。

因为特定链路的信息只产生于那些通过发送*HELLO*消息测试过链路状态的路由器，其他路由器只是毫无更改地重传此信息，所以与RIP路由器相比，过时信息不可能出现在OSPF路由器中。在OSPF路由器中，过时的信息很快被更新信息所取代，因为在链路状态改变之后，将会立即产生一条新信息。

OSPF网络中，也可能出现不稳定操作的时期。比如说，当有链路发生故障时，此信息不会立即到达所有路由器。如果此信息尚未抵达某一特定路由器，它会继续往目的网络发送分组，因为它认为链路是可用而且有效的。然而，这个时期并不会持续很长，分组也不会陷入路由循环。与之相反，如果不能通过故障链路传输分组，那么就简单地丢弃它。

OSPF协议的主要缺点是它的复杂性，随着网络规模的增长（即，随着网络数、路由数以及连接它们的链路的增长），它需要耗费的计算资源大量增加。为了克服这个缺陷，OSPF协议中引入了*领域*的概念，千万不要把它和因特网自治系统相混淆。属于特定领域的路由器只为该领域构建网络图，边界路由器只交换存在于每个领域中的网络地址信息以及边界路由器到各个网络的距离信息。在领域之间传输分组时，选择一个特定领域的边界路由器。通常，此路由器是距离所需网络最近的一个。

19.3.4　边界网关协议

如今，**边界网关协议版本4（BGPv4）**（border gateway protocol version 4）是因特网自治系统间交换路由信息的主要协议。开发BGP是为了取代EGP[⊖]，为因特网只有一个主干网时所使用。主干网是单个自治系统，其他自治系统根据树状的拓扑结构与之连接。因为这种结构消除了自治系统间循环的可能，EGP并没有采取任何额外的措施来防止路由循环的发生。

BGPv4成功运用于自治系统间的任意拓扑链路，后者与因特网的当前状态相对应。

让我们使用图19-18中的例子来解释一下BGP操作的主要规则。

在三个自治系统（1 021、363、520）的任意一个中，存在多个充当外部网关的路由器。这些路由器运行用来相互之间通信的BGPv4协议。只有当网络管理员在配置过程中明确指定这些路由器是它的BGP邻居时，路由器才会根据BGP与其他路由器相交互。举个例子，路由器EG1会根据BGP与路由器EG2相交互。这是因为在路由器EG1配置过程中，管理员指定路由器EG2的地址为194.200.30.2，是EG1的邻居，而不是因为这些路由器之间点对点相连。与之类似的是，配置路由器EG2的过程中，地址为194.200.30.1的EG1被指定为EG2的邻居。

在属于不同ISP的路由器交换路由信息时这种交互方法非常便捷实用，特定ISP的管理员或管理员们可能会决定ISP将使用哪个自治系统来交换流量，而使用哪个自治系统时不允许此类交换。这项任务通过为ISP外部网关配置邻居列表来实现。RIP和OSPF协议，设计为在自治系统内部使用，与它们可及范围内的所有路由器（使用LAN或点到点的链路）交换路由信息。这意味着所有网络的信息出现在每个路由器的路由表中，这样每个网络对于其他网络来说都是可及的。对于企业网络，这种情形很正常；对于ISP网络，却不希望这样。正因为这点，BGP在这里扮演着一个特殊的角色。

⊖ 这里的EGP是特殊路由协议的名称。回忆下，EGP这个简写同样可以作为自治系统之间路由整个协议分类的名称。这两个名字的重合造成了很多混淆。

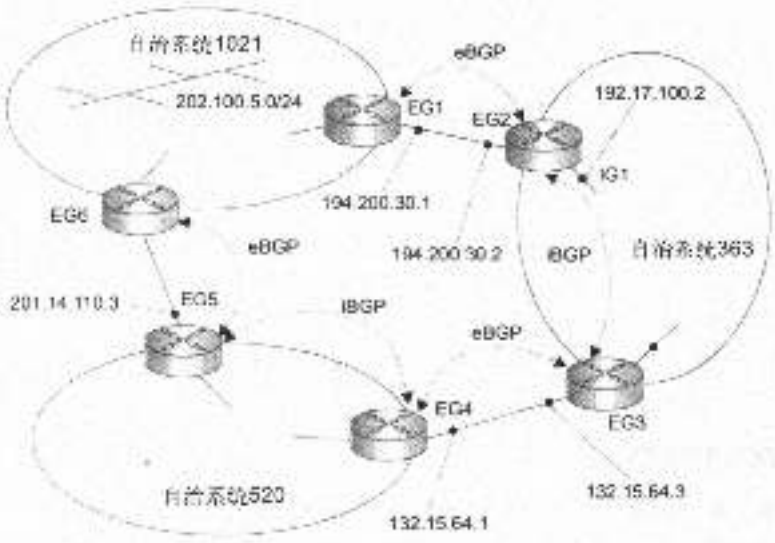


图19-18 使用BGP在自治系统间搜索路由

为了建立与特定邻居的连接，BGP路由器使用TCP协议（端口179）。在建立BGP连接时，可以使用不同方法的路由器认证来提升自治系统操作的安全性。

BGP的主要消息为更新（*update*）广告，利用此广告，路由器通知相邻自治系统的路由器它自己的自治系统网络是否可及。

该广告的名字说明了这些是只有自治系统中发生了变化才会发送给邻居的触发更新，这些变化可能与网络中新子网或路由的加入相关，或者可能是已经存在的子网或路由消失了。

在单个更新（*update*）消息中，可以通告单个新路由或者废除不再存在的多个路由。BGP将路由解释成必须被传递到路由中指定的各个子网的自治系统的一个序列。更加正式一点，BGP路由到子网的信息，在Network/Mask_length中指定，看起来如下：

BGP Route = AS_Path: NextHop: Network/Mask_length:

AS_Path表示自治系统号的集合，NextHop为路由器的IP地址，通过该路由器需要将分组传到Network/Mask_length网络。打个比方说，如果路由器EG1需要通知路由器EG2一个新子网，202.100.5.0/24，已经出现在自治系统AS1021之中，它会生成如下消息：

AS1021; 194.200.30.1; 202.100.5.0/24,

然后，路由器将它传递到自治系统AS363的路由器EG2。很自然，在如此做之前路由器EG1必须建立到路由器EG2的BGP连接。

路由器EG2，在收到更新消息之后，在路由表中保存携带NextHop 194.200.30.1子网202.100.5.0/24相关的信息。同时标记出此消息来自于BGP。路由器EG2与AS363的内部路由器利用IGP组中的某一协议来交换路由信息，如，可以是OSPF。如果EG2路由器被配置为在BGP路由到OSPF路由的重分配模式下操作，那么AS363所有的内部路由将会从外部的OSPF广告中得知202.100.5.0/24网络的存在。现在，EG2路由器将它自己的内部接口地址指定为NextHop，如，可能是192.17.100.2（对于IG1）。

然而，为了将与202.100.5.0/24网络相关的广告传播到其他自治系统，比如AS520，不能使用OSPF协议。连接到自治系统AS520中路由器EG4的路由器EG3必须使用BGP来生成所需格式的更新（*update*）消息，为了执行此项任务，它不能使用通过内部接口所接收到来自OSPF的与202.100.5.0/24相关信息，因为它格式不同而且不包含自治系统号的信息。

为了解决这个问题，EG2和EG3必须也利用BGP建立一个会话，尽管它们属于同一个自治系统。与它的主要用法外部BGP相比，BGP的这种用法称为内部BGP。因此，路由器EG3从路由器EG2接收所需信息并且传送给它的外部邻居，路由器EG4。当生成新的更新消息时，EG3通过将自己的自治系统，A520，加入到自治系统列表以传递来自于路由器EG2的消息并且用它自己接口地址来取代收到的NextHop值：

AS363, AS1021; 132.15.64.3; 202.100.5.0/24.

自治系统号消除了更新消息循环。比如说，当路由器EG5传递有关网络202.100.5.0/24的消息给路由器EG6时，后者并不会使用，因为此消息看起来如下：

AS520, AS363, AS1021; 202.14.110.3; 202.100.5.0/24.

因为自治系统列表已经包含了本地自治系统号，很明显该消息是循环的。

如今，BGP的用途远远超过了在自治系统间交换路由信息。

19.4　因特网控制报文协议

ICMP在网络中扮演着辅助角色，它的详细定义在RFC 792中描述。

有些情况下IP不能向目的主机发送分组。例如，当分组的TTL已经过期时，当到达指定目的地址的路由从路由表中丢失时，当分组没有通过校验和的验证时，或者当网关没有足够的缓存空间来传递特定分组时，都有可能会发生。正如我们已经提到的，IP根据**尽力服务**（best effort）原则操作，意味着它并不采取任何措施来确保数据的传输。IP的这个特征由更高层协议，如传输层的TCP或某种程度上被应用层的DNS来弥补。这些协议承担了确保可靠性的职责。它们使用了诸如消息编号、传送确认和数据重传等有名的技术来实现此目的。

ICMP对IP的功能进行补充；但是，这补充的本质与由高层协议确保可靠传输的能力不同。ICMP并不旨在解决分组传输过程可能出现的问题：如果分组丢失了，ICMP并不能进行重传。ICMP的目标更加简单。该协议负责通知发送端有关分组发生了“意外”的信息。因为IP发送分组随后就“忘记”了它的存在，ICMP“跟踪”分组在网络上传输的过程；如果路由器丢弃了它，ICMP发送消息给源主机以通知它这个事件的发生。因此，ICMP保证了被发送的分组和发送端主机之间的持续反馈。

举个例子来说，假定运行在特定路由器上的IP察觉到沿着路由传送的分组需要分片而它的不准分片（*DF*）标志却被设为1。IP模块发现它不能继续传送此分组，必须发送ICMP诊断（*diagnostic*）消息给源主机，然后丢弃掉该分组。

除了诊断，ICMP还被用来监督网络。比如说，IP网络流行的诊断工具ping和tracert利用ICMP报文来工作。利用ICMP报文，应用程序可以判断数据沿着哪条路由传送、评价它的可用性、断定数据送到特定主机所需的时间、为特定网络接口申请掩码值，等等。

注意有些分组会在没有确认的情况下从网络上消失。特别地，ICMP并不传递有关在处理携带ICMP错误消息的IP分组过程中产生问题的消息。（但是，这个规则不适用于ICMP请求。）该协议的开发者利用这种解决方案来避免网络中“风暴”的产生，即错误消息数量急剧膨胀。出于同样的目的，如果错误发生在除了第一个分片以外的任何分片的传输过程中，如果丢失的分片拥有广播IP地址，或者如果丢失的分组被封装在低一级技术并携带广播地址的帧中，也不会发送ICMP报文。

因为IP分组包含发送端地址而不包含任意中继路由器的地址信息，所以ICMP报文只被发送到终端节点。在这里，操作系统核心可以通过传输层或应用层协议，或者通过应用程序，来处理这些消息。另一方面，也可以忽略这类消息。最重要的一点就是ICMP报文的处理不属于IP和ICMP的责任。

19.4.1 ICMP报文的类型

所有的ICMP报文归属于以下两类：

- 诊断（错误）报文
- 信息报文比如请求和响应

ICMP报文被封装在IP分组的数据字段中（图19-19），一个ICMP头为8字节长并包含如下字段：

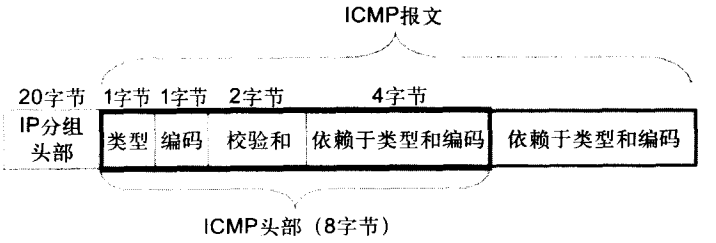


图19-19 ICMP报文格式和封装

类型（type）（1字节）——包含判断报文类型的代码。表19-7列出了最常用的ICMP报文类型。

表19-7 类型字段的取值

取值	报文类型	取值	报文类型	取值	报文类型
0	回送应答	8	回送请求	14	时间戳应答
3	目的地不可达	11	数据报超时	17	地址掩码请求
4	源抑制	12	数据报的参数问题	18	地址掩码应答
5	重定向	13	时间戳请求		

代码（code）（1字节）——包含更加准确区分错误类型的代码

校验和（checksum）（2字节）——为整个ICMP报文计算的校验和字段

头部还包含4字节组成的一个字段，该字段的内容依赖于类型（*type*）和代码（*code*）字段的值。在请求/响应报文中，该字段包含2字节的标识符（*identifier*）以及序列号（*sequence number*）子字段（图19-20）。包含在这些子字段中的号码从请求报文复制到响应报文。标识符字段允许目的主机来判断该响应是发给哪个应用程序的，序列号（*sequence number*）字段是应用程序用来识别与特定请求相关的响应（考虑到同样应用程序会产生多个同种类型的请求）。在错误报文中，并不使用此字段，因此填充全零。

各种类型的错误都具有错误代码。比如说，表19-8提供了类型3报文的代码——“目的地不可达”。这张表列出了15个可以在这类报文中指定的理由。不能抵达目标主机可能是因为，比方说，临时硬件故障、不正确的目的地址、缺乏应用层协议或目标主机缺乏开放的UDP/TCP端口等等。

ICMP数据字段的格式也依赖于类型（*type*）和代码（*code*）字段的值。为了证实不同报文类型格式上的区别，请考虑下面两个例子：

- 回送请求/响应
- 目的地不可达

表19-8　详细说明类型3错误——“目的地不可达”的代码

代码	原　因	代码	原　因	代码	原　因
0	网络不可达	6	目的网络未知	12	由于ToS主机不可达
1	主机不可达	7	目的主机未知	13	通信被管理员过滤后禁止
2	协议不可达	8	源主机孤立	14	主机优先权冲突
3	端口不可达	9	目的网络被管理员禁止	15	免除优先权
4	需要分片，但是设了DF位	10	目的主机被管理员禁止		
5	源路由失败	11	由于ToS网络不可达		

19.4.2　回送请求/响应报文的格式：Ping实用程序

图19-20说明了“回送请求”和“回送响应”报文的格式。它们之间的区别仅仅在类型字段的值（0是响应，1是请求）。在请求数据字段中，发送端填充信息，随后它又从目的主机的响应中收到此信息。

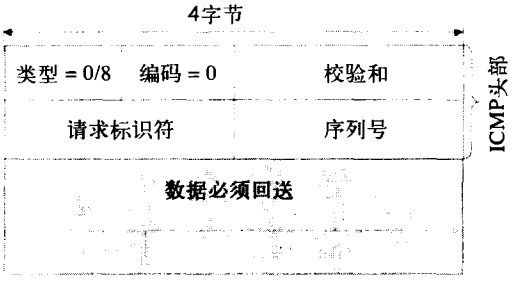


图19-20　回送请求/回送响应ICMP报文的格式

回送请求和回送响应，集合起来叫做**回送协议**（echo protocol），是网络监控最简单的工具。计算机或路由器通过因特网发送一条回送请求，其中指定了主机的IP地址，虽然需要检查该IP是否可达。收到回送请求的主机，生成并发送一个回送响应，然后将报文返回给发送请求的主机。因为回送请求和回送响应都以IP分组的形式在网上传递，所以它们的成功传送意味着整个传输系统都正常工作。

大多数操作系统使用内嵌的ping工具，用于测试抵达主机的可能性。通常，此工具发送一系列回送请求给被测试的主机并且向用户提供丢失回送响应以及网络响应请求平均时间的统计信息。Ping工具显示报文以通知用户所有收到的响应。输出看起来大致如下：

```
# ping server1.citmgu.ru
Pinging server1.citmgu.ru [193.107.2.200] with 64 bytes of data:

Reply from 193.107.2.200: bytes=64 time=256ms TTL=123
Reply from 193.107.2.200: bytes=64 time=310ms TTL=123
Reply from 193.107.2.200: bytes=64 time=260ms TTL=123
Reply from 193.107.2.200: bytes=64 time=146ms TTL=123
```

从这里提供的列表中，可以看出作为发送到主机server1.citmgu.ru的测试请求，结果是收到四个回送响应。每条消息的长度为64字节，下一列包含RTT的值，从请求发送时刻起到收到响应的时刻之间的间隔。正如你所看到的，网络操作是不稳定的，因为最后一行的RTT值比第二行的少两倍多。到达分组的TTL值显示在第三列中。

依赖于ping工具的特定实现以及命令行选项，命令行输出可能与这里显示的有所区别。通常，ping工具有多个命令行选项可以用来指定消息数据字段的大小、TTL字段的初始值、分组传送重复尝试的次数以及DF标志的设置。

如果在指定超时期间内没有响应抵达，或者如果ICMP响应带有错误消息，ping会在屏幕上显示适当的错误消息。

19.4.3　错误报文格式：Traceroute实用程序

图19-21显示了ICMP错误报文的格式；在此例中，这就是“目的地不可达”报文。其他ICMP

错误报文与之格式相同只是在类型（*type*）和代码（*code*）字段的值上有所区别。

如果路由器不能发送或传输IP分组，它发送“目的地不可达”的错误报文给刚发送此分组的主机。在该报文中，类型（*type*）字段值为3，代码字段填充了范围为0到15的号，此号更精确地指明了分组不能传送的原因。紧跟着该字段的4字节是校验和（checksum）字段，在这里未使用因而填充为全零。

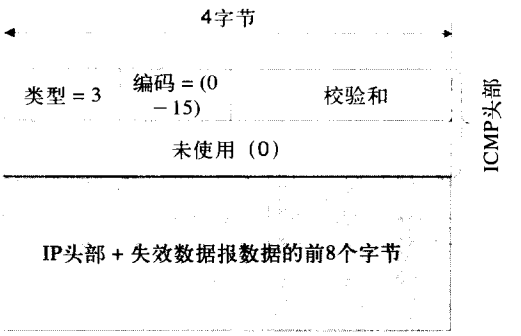


图19-21 ICMP错误报文“目的地不可达”的格式

除了ICMP头部指明的错误原因，ICMP总是用IP头部和引起错误的IP分组数据字段的前8个字节来填充它的数据字段。此信息允许发送端更加精确地断定错误原因，因为TCP/IP栈利用IP分组传输消息的所有协议都将最重要的信息放在报文的头8个字节中。特别地，这些可能是TCP或UDP头部的头8个字节，其中包含识别发送了丢失分组的应用程序的信息。因此，应用程序开发者可以提供用于响应不能发送分组的ICMP消息的内嵌工具。

目标主机或目标网络可能因为临时硬件故障、发送端指定的错误目标网络地址或路由器没有到目标网络的路由信息，而变得不可达。不能抵达协议、端口或两者皆不可达意味着目标主机缺乏某些应用层协议的实施或目标主机缺乏开放的UDP或TCP端口。

正如在ping工具示例中已经说明的，ICMP报文用于网络监控时很有效。特别是，另外一个错误“超时”的错误报文用作为另外一个常用工具的基础——traceroute（UNIX）或tracert（Windows 2000）。该工具允许跟踪到远程主机的路由，为每个中继路由器定义RTT、IP地址以及域名（假设这名字在DNS反向搜索区域中注册过）。该信息对于定位分组到目标主机的路径突然崩溃的路由器很有用。

Traceroute通过发送目的地址为路由终点的普通IP分组来跟踪路由，跟踪方法的思想在于第一个发送的分组TTL设为1。在第一个路由器的IP接收到此分组之后，它根据算法将TTL减1。TTL值变为0，路由器丢弃TTL为0的分组并返回携带IP头部和丢失分组的首8个字节的“超时”报文给源主机。

在收到ICMP报文通知发送端分组未发送的原因之后，traceroute工具记忆下第一个路由器的地址（从携带ICMP报文分组的IP头部中检索出）。接着，它计算第一个路由器的RTT值。然后，traceroute发送下一个TTL值设为2的IP分组。该分组成功路过第一个路由器但是会被第二个丢弃，于是立即发送同样的“超时”ICMP报文。Traceroute工具存储下IP地址和第二个路由器的TTL值，诸如此类。同样的操作在到目的主机的路由途中每下一个路由器上执行。

显然，我们考虑了traceroute工具最简单形式的操作。可是，即使是这个信息也已经足以评价出它所基于的思想是多么地优雅！

下面提供的列表显示了tracert工具（Windows）在跟踪ds.internic.net [198.49.45.29] 主机过程中的典型输出：

```
1  311 ms  290 ms  261 ms  144.206.192.100
2  281 ms  300 ms  271 ms  194.85.73.5
3  023 ms  290 ms  311 ms  Moscow-m9-2-S5.relcom.eu.net [193.124.254.37]
4  290 ms  261 ms  280 ms  MSK-M9-13.Relcom.EU.net [193.125.15.13]
5  270 ms  281 ms  290 ms  MSK-RAIL-1-ATM0-155Mb.Relcom.EU.net [193.124.254.82]
6  300 ms  311 ms  290 ms  SPB-RASCOM-1-E3-1-34Mb.Relcom.EU.net [193.124.254.78]
7  311 ms  300 ms  300 ms  Hssill-0.GW1.STK2.ALTER.NET [146.188.33.125]
8  311 ms  330 ms  291 ms  421.ATM6-0-0.CR2.STK2.Alter.Net [146.188.5.73]
```

```
9  360 ms   331 ms   330 ms  219.Hssi4-0.CR2.LND1.Alter.Net [146.188.2.213]
10 351 ms   330 ms   331 ms  412.Atm5-0.BR1.LND1.Alter.net [146.188.3.205]
11 420 ms   461 ms   420 ms  167.ATM8-0-0.CR1.ATL1.Alter.Net [137.39.69.182]
12 461 ms   441 ms   440 ms  311.ATM12-0-0.BR1.ATL1.Alter.Net [137.39.21.73]
13 451 ms   410 ms   431 ms  atlanta1-br1-bbnplanet.net [4.0.2.141]
14 420 ms   411 ms   410 ms  vienna1-br2.bbnplanet.net [4.0.3.154]
15 411 ms   430 ms  2514 ms  vienna1-nbr3.bbnplanet.net [4.0.3.150]
16 430 ms    421 ms   441 ms  vienna1-nbr2.bbnplanet.net [4.0.5.45]
17 431 ms    451 ms   420 ms  cambridge1-br1.bbnplanet.net [4.0.5.42]
18 450 ms    461 ms   441 ms  cambridge1-cr14.bbnplanet.net [4.0.3.94]
19 451 ms    461 ms   460 ms  attbcstoll.bbnplanet.net [206.34.99.38]
20 501 ms    460 ms   481 ms  shutdown.ds.internic.net [198.49.45.29]
Tracing complete.
```

以上行的顺序对应着形成到目的主机路由的路由器顺序，每行的第一个数字是相应路由器的跳数。Traceroute工具测试每个路由器三遍；因此，接下来的三列是TTL在那个路由器上过期所发送的三个分组算出的RTT值，如果来自特定路由器的响应并没有在指定时间内到达，那么会在时间的地方显示一个星号（*）。

然后，指定了IP地址和域名（如果可获得的话）。从这个列表中，很显然，事实上所有ISP的路由器接口都在DNS服务中注册过；前面两个对应着本地路由器，没有注册。

需要再次指出的是每行中指定的时间并不是分组在两个邻居路由器间传送所需的时间。这是分组从源传输到恰当路由器再返回的时间。因为因特网路由器的负载一直在变化，传输到特定路由器的时间并不总是在稳定地增加。有时候，它会任意变化。

小结

- 与IP相比，其主要任务是确保互联网上网络接口之间数据的传输，TCP和UDP的主要目标为在网络不同终端节点上的应用程序进程之间传输数据。
- TCP与UDP的主要区别在于TCP执行附加任务——确保互联网上信息的可靠传输，其中所有节点使用一种数据不可靠传输的IP数据报协议。
- UDP是根据尽力服务原则操作的一种数据报协议，它不建立逻辑连接。UDP并不保证报文传送，因此，不能补偿同样是数据报协议的IP的不可靠性。
- 应用程序进程访问点的系统队列叫做端口。端口根据编号以及计算机范围内唯一定义的应用程序来识别。使用UDP的应用程序拥有称为UDP端口的编号，依赖TCP的应用程序则拥有TCP端口。
- 如果应用程序进程为常用的公共服务，比如FTP、telnet、HTTP、TFTP或DNS，它们拥有统一分配的端口号。不广泛使用以及没有分配知名端口号的服务拥有本地操作系统分配的端口号。这种端口号被称为动态的。
- 应用程序的套接字为下列参数对：（*IP地址，端口号*）。
- TCP建立逻辑连接来解决确保可靠数据交换的问题。逻辑连接由套接字对来唯一标识。
- TCP连接是一个全双工连接；它建立在对下列参数协商的结果：最大MTU大小；在没有收到确认的条件下可传输的最大数据容量，从特定连接范围内字节流起始的字节序列号开始算起。建立了连接之后，两端的操作系统都为之分配一套系统资源。这些资源被用来组织缓存、计时器以及定时器。
- TCP/UDP对多个应用程序服务发来的数据进行处理的过程叫做多路复用。TCP/UDP用来将网络层的分组扩散到更高层服务的相反过程叫做解多路复用。UDP在套接字基础上实施解多

路复用，TCP在连接的基础上执行同样的任务。

- 为了控制TCP连接框架中的流量，使用了一种叫做滑动窗口协议的特殊方法。接收端向发送端传送以字节为单位的窗口大小。然而，发送端也能够控制窗口的大小。如果发送端注意到通信链路的操作是不可靠的，它可以将窗口大小缩减到初始值。
- 为了各个路由器之间协调一致的路由表而产生了路由协议——也就是，在有限步数内确保沿着一个合理的路由传送分组。为了实现此目标，网络路由器彼此交换有关因特网拓扑的信息。
- 路由被分为下面两大类：静态的和自适应的（动态的）。
  - 在静态路由过程中，所有路由表由网络管理员手动编写并放入到每个路由器中。
  - 自适应路由在网络配置改变之后确保路由表的动态更新。
- 自适性路由协议被分为两组，每个与特定类型的算法相关联：
  - 在距离向量算法（DVA）中，每个路由器阶段性地在网络上发送广播报文，其中含有表示从这个路由器到所有已知网络的距离。
  - 链路状态算法（LSA），向每个路由器提供足够构建准确网络链路图的信息。
- 因特网路由协议被分为内部的和外部的两种。外部网关协议（EGP）在自治系统之间传递路由信息，内部网关协议（IGP）只用于特定的自治系统内部。
- RIP是TCP/IP网络中最古老的路由协议。虽然它很简单，因为使用DVA所造成，但RIP成功用于不超过15个中继路由器的小型网络中。
- RIP路由器通常基于最简单的度量来选择路由，仅仅根据网络中的中继路由器数，即跳数来判断。
- 在使用RIP并存在循环路由的网络中，当分组陷入循环而不能传送到目的地时，可能会出现相当长的不稳定操作时期。RIP路由器提供了偶尔会减少不稳定时期的多项技术。这些技术包括水平分割、压缩以及触发更新。
- 在选择路由时，OSPF路由器使用考虑了网络带宽的度量。
- OSPF协议允许在路由表中存储到同一网络的多个路由，条件是它们拥有相等的度量。这使得路由器可以在负载平衡状态下操作。
- OSPF相当复杂并且需要很强的计算能力；因此，它通常在强劲的硬件路由器上操作。
- 如今，边界网关协议（BGP）版本4是在因特网自治系统之间交换路由信息的协议。BGPv4具备高度的可靠性，拥有任何与当前因特网结构所对应的自治系统之间的链路拓扑。
- 因特网控制报文协议（ICMP）扮演着网络上的辅助角色。它被用于诊断以及网络监控。因此，ICMP报文成为监控IP网络常用工具，如ping和tracert的基础。

复习题

1. 什么时候软件开发者更倾向于使用UDP？什么时候它们倾向于依赖TCP？
2. 在TCP会话过程中，TCP分段的发送端会收到多少容量的数据（精确到1个字节）？TCP分段头部的ACK字段值为1 845 685，已知收到的第一个字节号为50 046。
3. 当路由器缺少路由表时还可以传递IP分组吗？

   A. 不，不可能。

   B. 可以，假设使用源路由就可以。

   C. 可以，如果路由器中指定了默认路由就可以。
4. 网络中会出现缺乏路由协议的情况吗？

5. 距离向量路由协议的缺点是什么？
   A. 大型网络中巨大的流量
   B. 所选择的路由经常不具备最小度量
   C. 协调路由表的过程耗费大量时间
6. 基于LSA的路由协议主要操作原则是什么？
7. IGP与EGP之间区别何在？
8. RIP中使用了哪种度量？
9. 为什么RIP认为16跳的距离是不可达的？
   A. 分配给距离值的字段长度为4个二进制数字
   B. RIP算法操作的网络很少那么大
   C. RIP企图确保一个可接受的算法收敛时间
10. 加速RIP收敛速度的方法有哪些？
11. 使用OSPF构建路由表的主要步骤是什么？
12. OSPF协议中*HELLO*消息扮演的是什么角色？
    A. 它们在两个路由器之间建立连接
    B. 它们检查通信链路和邻居路由器的状态
    C. 它们携带OSPF协议在网络中操作的信息
13. OSPF支持哪种类型的度量？
14. 支持OSPF的路由器网络被分割成领域的目的何在？
15. OSPF的主要缺点是什么？
16. 为什么EGP不再在因特网中使用？
17. 允许EGP在自治系统之间存在循环的网络中操作的机制是什么？
18. 从一个自治系统收到广告之后，BGP路由器在将它传递到下一个自治系统时会修改什么参数？
19. 当IP分组出现问题时，什么情况下不能发送ICMP错误报文？
20. ICMP报文的目的地是什么？哪个软件模块负责处理它？
21. ICMP报文如何提高IP网络中数据传输的可靠性？

练习题

1. 找个伙伴一起建立一个TCP连接。讨论最大分段尺寸、缓存的初始值、序列号的初始值以及窗口大小。然后，相互间异步地发送“分段”。“分段”的角色可以由卡片充当，上面写着关键字段的值——首字节号、发送分组的大小、确认号以及窗口的新值。你可以假装偶尔丢失了卡片并根据TCP操作逻辑来处理。不要忘记为每个发送的分段标上时间来跟踪确认是否到达。这个游戏会帮助你更好地理解TCP。过程中尽量相互提问。
2. 图19-15中所示的网络中，在路由器R1丢失了到网络201.36.14.0的连接后，要想多个路由表重新达到协调状态，在最坏的情况下，需要多久？假设所有路由器支持水平分割机制。
3. 说出几个同时考虑到带宽、可靠性、通信链路延迟的度量。

第20章 IP路由器的高级特性

20.1 引言

IP路由的主要功能是建立路由表并在这些表的基础上传递IP分组。为了执行这些功能，路由器必须支持第18章中描述的因特网协议（IP）以及第19章中提到的路由协议。除了这些基本功能以外，如今的IP路由器还支持多个高级特性，这使得它们的功能更加强大并能更好地处理各种各样的多功能流量。本章中，我们将考虑当今路由器最重要的高级特性，也是网络管理员最常用到的。

路由器是连接计算机网络与外部世界的边界设备。因此，由它们来保护内部网不受外部侵袭是理所当然的了。IP通过用户流量过滤来执行这些功能，过滤根据IP分组的不同属性来进行，比如源地址、IP分组中封装的协议类型以及产生该流量的应用程序。这种功能阻止了不想要的流量流入受保护的网络，同时减少了网络内部主机受到攻击的可能性。网络地址转换（*NAT*）技术，向外部用户隐藏网络中主机的真实地址，在保护这些内部网络资源方面扮演着重要的角色。

*QoS*支持是IP网络中相对较新的特征。IP路由器一直以来支持多种拥塞控制及避免的机制。但是，一直到20世纪90年代后期，确保IP网络的QoS支持的标准最近才有所发展。IP网络有两种QoS架构：**集成服务**（IntServ）和**区分服务**（DiffServ）。第一种架构保证了单个流量从源到目的地的QoS，第二种为聚合流量而开发，代表着少量的流量分类。如今，集成服务技术主要用在网络边界设备、企业范围的网络以及访问网络中。区分服务则开始运用在主干网络上。应用领域的区分是很明显的，因为确保单个流量的QoS在路由器上造成了额外的负担，此负担与服务的流量数成一定比例。主干网可以传输几十万的用户流量；因此，集成服务的实施可能会对主干路由器的计算能力和内存空间提出过高的要求。

本章最后会探讨当前路由器的功能性结构。

20.2 过滤

IP路由协议创建路由表。基于此，互联网的任何主机都可以和其他主机交换信息。数据报网络的原则通常是让人愉快的，因此，因特网的每个用户都可以访问任何公共站点。

回想一下在基于虚拟循环技术的网络中，在任何两个主机之间若没有事先建立虚拟电路的话就不可能通信。

但是，主机和网络这样共同的可用性并不总是与拥有者的需求相对应。因此，很多路由器支持高级特性，比如用户流量过滤和路由协议广告。这些能力将主机可达性控制在不同等级的粒度上。

20.2.1 用户流量过滤

过滤（filtering）是路由器对IP分组的非标准处理，会导致特定分组的丢弃或者路由的改变。

路由器实施的用户流量过滤，原则上类似于LAN交换执行的功能（参见第15章）。

过滤的主要目的在于并不是所有经过路由器的IP分组都必须根据第18章中描述的标准过程来处理。在标准过程中，分组处理的类型是根据IP分组[⊖]目的地址字段中包含的信息而选择。如果这

⊖ 此描述被简化了以强调传递分组标准过程的主要特性。事实上，即使标准过程考虑了IP分组的多个其他字段，如TTL、数据字段（名义上，是它的大小）、DF、IP优先级以及TOS。如果这些字段不需要分组进行特殊处理（比如当TTL<1是丢弃分组），那么路由器就启动查找路由表的进程。

个地址存在于路由表中，那么该分组就被传送到恰当的输出接口。

与LAN交换实施的类似过滤比起来，路由器上设置的分组过滤条件通常考虑到更多的属性。比如说，除了目的IP地址以外，这种属性的列表可能还包括：

- 源和目的IP地址
- 源和目的MAC地址
- 所收到分组的接口标识符
- IP分组中封装的协议类型（比如，TCP、UDP、ICMP或OSPF）
- TCP/UDP端口号（如，应用层协议的类型）

如果设了过滤，路由器首先会检查过滤的条件是否满足。如果检查结果为肯定的，那么路由器为该分组执行一些非标准的操作。比如，分组可能被抛弃（丢弃）、发送到下一个路由器（与路由表中指定的不同）、或者在网络拥塞发生时将它标记为一个丢弃的候选。根据路由表记录传递分组的正常过程可能是其中一个操作。

现在来看几个用Cisco IOS命令行接口语言写的过滤示例，这些过滤也叫做**访问列表（access lists）**，被广泛用于限制IP路由器中的用户流量。

标准访问列表（standard access list）是最简单的过滤，只考虑了源IP地址。

这种条件的通常语法看起来如下：

访问-列表-访问-列表-号{拒绝 | 允许}

{源-地址 [源-模糊匹配] | 任意}

标准访问列表定义了当分组满足过滤所描述的条件时，分组执行的两种动作：*拒绝（deny）*（即丢弃分组）以及*允许（permit）*（即根据路由表传递分组以供标准化处理）。当源IP地址匹配列表中指定的源IP地址时，就选择标准访问列表中的对应动作。这种检查的执行过程与检查路由表时相同。这里，*源-模糊匹配（source-wildcard）*类比于掩码——然而，有一点小小的改变。源-模糊匹配中的二进制0意味着到达分组地址中的这一位必须匹配过滤条件中指定的地址，二进制1意味着此位置上的匹配是不需要的。如果你需要为特定子网的所有地址指定一个条件，你必须使用该子网掩码的相反值。*任意（any）*参数的意思是地址的任何值；它只是一个简写并且等同于为*源-模糊匹配*（source-wildcard）字段指定255.255.255.255值，这样更容易理解。

标准访问列表的示例如下：

访问-列表 1 拒绝　192.78.46.0　0.0.0.255

这里：

1——访问列表号

拒绝——对满足访问列表中指定条件的分组必须执行的动作

192.78.46.0——源地址

0.0.0.255——源模糊匹配

该过滤阻碍了以下分组的传送，即地址中三个最高位的值为192、78和46的分组。

访问列表可以包含多个条件。这时，它包含多个起始于*访问-列表（access-list）*关键字的行，行数等于代表列表标识符的*访问-列表-号（access-list-number）*的值。举个例子，如果你需要允许源于主机192.78.46.12的分组传递到路由器，不允许属于该子网的其他主机传输该分组，那么访问列表看起来如下：

访问-列表 1 允许 192.78.46.12　0.0.0.0

访问-列表 1 拒绝 192.78.46.0　0.0.0.255

访问-列表 1 允许 任意

挨个检查访问列表条件。当任意一个匹配成功时，则执行条件中指定的*允许（permit）*或拒

绝（*deny*）动作。然后，就不再检查剩余的条件。默认条件下，在每个列表的末端，都有下列隐含条件：

[访问-列表 1 拒绝 任意]

因此，为了防止列表干涉来自其他网络分组的正常处理，往其中写入下列条件：

访问-列表 1 允许 任意

访问列表可以被应用于路由器的任何接口、任何方向。如果列表与*in*关键字一起出现，它应用于输入分组；如果是关键字*out*，该条件应用于输出分组。举个例子，访问-列表1可以利用下列命令应用于输入流量的某一接口：

访问-组 1 in

Cisco路由器拥有更多强大的访问列表类型，比如扩展访问列表。这些列表常用的格式如下：

访问-列表 访问-列表-号 {拒绝 | 允许}

{协议 | 协议关键字}

{源地址 [源-模糊匹配] [源端口] | 任意}

[目标地址 [目标-模糊匹配]] [目标端口]

使用扩展访问列表，可以阻止使用TCP 21号端口接收客户请求的ftp流量传入公司的内部网。为了达到此目标，需要在访问列表中包含下列条件：

访问-列表 102 拒绝 TCP 任意 21 任意

这必须应用于关键字为out的内部网所连接的路由器接口。

企业范围网络的管理员通常利用下列条件来禁止ping内部主机：

访问-列表 101 拒绝 ICMP 任意 192.78.46.8 0.0.0.0 eq 8

从这个条件中可以看出，ICMP的访问-列表语法与用于扩展访问列表的标准语法不同。参数8意味着不允许传递类型为8的ICMP消息，它对应于ping工具使用回应-请求消息。

在很多UNIX版本中，gated软件路由器使用的过滤语言更加灵活多变。此语言使用一种类C语言的语法，允许使用if、then和else逻辑操作符来构建相当复杂的逻辑操作。

需要提到的是，用户流量的过滤会大大降低路由器的操作，因为处理每个分组时都需要检查辅助的条件。

为了避免路由器严重过载，从而使它疏于执行其主要任务，路由器过滤并不使用会话建立前的信息。不论过滤条件有多复杂，它只包含当前分组的参数并且不能使用路由器已经处理分组的参数。这个限制是路由器与防火墙的主要区别，后者是使用会话前信息执行更好过滤的特殊软件系统。

20.2.2 路由公告过滤

为了控制到达特定主机和网络的可能性，可以限制与用户流量过滤一起的路由公告的繁殖。

此方法阻止某网络上记录自动出现在路由表中。因为到达路由器的路由广告远比用户分组少，这个方法大大削减了路由器的能力。

Cisco路由器提供了限制特定网络上路由广告繁殖的可能性，它使用一个标准访问列表来申明描述，然后使用分布-列表（*distribute-list*）关键字（取代了过滤用户流量中使用的访问-组关键字）将之应用于接口上。

举个例子，如果网络管理员想要阻止公司内部网194.12.34.0/24和132.7.0.0/16上的信息繁殖扩散到外部网络，定义如下标准访问列表就足够了：

访问-列表 2 拒绝 194.12.34.0 0.0.0.255

访问-列表 2 拒绝 132.7.0.0 0.0.255.255

访问-列表 2 允许 任意

然后，管理员使用下列命令将之应用于接口：

分布-列表 2 out 系列 1

20.3 IP QoS

TCP/IP栈技术为了灵活流量而创建，它能够容忍分组延迟以及分组延迟的各种变化形式。因此，TCP/IP开发者的注意力一直集中在确保TCP的可靠流量传输上。然而，为了减少过载并防止缓慢链路的拥塞，很多QoS机制逐渐地进驻IP路由器，包括优先级和权重队列、流量策略以及反馈。每个网络管理员都随心所欲使用这些机制而不用考虑系统整体的解决方法。到了20世纪90年代中期，才开始发展IP QoS标准领域的研究。在这些标准基础上，在互联网框架内甚至互联网内创建一个QoS支持系统才成为可能。

作为研究的结果，发展了两个IP QoS系统：

- 集成服务（*integrated service*，*IntServ*）为终端用户数据流提供QoS保证。因此，IntServ主要用在网络外围设备上。
- 区分服务（*differentiated service*，*DiffServ*）为流量分类做同样的事。因此，主要用于主干网上。

两个系统使用基于资源保留的QoS系统的所有基本元素——即，两个系统都提供下列机制：

- 为流量加设条件
- 路由器协调的信令
- 为流或流量分类保留接口和路由器带宽
- 优先级和加权队列

因为分组依然沿着最佳度量值的路径转发，利用标准路由协议进行选择而不考虑通信链路的真实负载，所以这些技术都不能解决流量工程问题。

20.3.1 IntServ和Difdserv QoS模型

IntServ技术的发展在20世纪90年代由IETF发起。这是第一个发展和研究领域，在其中利用系统方法解决了确保TCP/IP网络QoS的问题。Basg集成服务模型假定网络路由器的集成通信以确保网络终端节点间沿着**微小流完整路由**（entire route of the microflow）所需的QoS。

路由器资源（比如，接口带宽和缓存大小）根据不同应用程序的QoS需求分布在给定网络QoS策略允许的范围内。QoS应用程序请求使用RSVP指示协议在网络上扩散，它允许为两个终端节点间流量（如单播目标地址）以及多个终端节点收到的流量（如多播目的地址）保留资源。

但是，这个确保QoS的和谐系统有无数的反对者，主要是ISP。这是因为实施IntServ需要ISP的主干路由器处理途径ISP网络几十万个微流的状态信息。路由器上这种非传统负载需要重新设计它们的结构，这大大增加了IP网络和服务的费用。

因此，到了20世纪90年代后期，开发了另外一种IP QoS技术，叫做**区分服务**（differenciated service，DiffServ）。起初，这个技术面向在ISP网络范围内使用，它排除了产生计划外微流的节点。对于区分服务，QoS支持始于IPS网络的边界路由器，大量来自用户网络的微流到达该路由器。每个区分服务的边界路由器对输入流量进行分类并打上标签，将它们划分为几个类别，通常不超过3～4（最高为8）。然后，每个网络路由器，通过分配一定量的资源给每个类别，在区分基础上根据打上的标签来为各个流量类别服务。路由器上的资源是静态保留的。最常见的情况下，这个任务由网络管理员手动执行。信令协议的角色由标签来扮演，这些标签指定特定分组属于某个分类。

管理员负责所有网络路由器之间流量的协调服务，因为管理员决定带宽以及负责给每个路由器的每个接口上的每个流量分类分配缓存空间。

区分服务模型大大减轻了ISP路由器的负载，因为它只需要为少量流量分类存储状态信息。而且，这个模型对于ISP来说也受欢迎，因为它们可以在ISP所拥有的网络范围内组织自动的QoS支持。但是，这些好处是要付出代价的。在这里，获得这些好处必须要付出端对端QoS支持的代价，即使每个ISP在自己的网络中实施区分服务，整个模式也是分块的，因为单个管理员只负责每一小块。保留参数的协调仍然尚未被考虑到，也不被任何协议所支持。

尽管最近对区分服务的大量关注，认为它是一个能够用很小的代价来确保更多因特网QoS的简单工具，但仍然存在着不同的观点。比如说，Lawrence G. Roberts博士，因特网创建人之一，对于尝试以简单的形式来解决因特网QoS问题，表达了强烈的反对意见。集成服务和区别服务技术联合使用的研究也正在进行中，在这些模型中，每个技术在其应用领域里操作：集成服务应用在访问网中，微流的数目相对较小，而区别服务则应用在主干网中。另外一个替补区分服务的成员是多协议标记交换（MPLS）技术，它允许在IP网中解决流量工程问题。该技术将在第五部分中更加详细地介绍，第五部分专注于WAN技术，因为该技术通过将IP与流行的WAN技术如ATM相捆绑而得以出现。因此，在学习了ATM之后再研究会显得容易很多。

集成服务和区分服务（RFC 3290）依赖于同样的基本QoS机制。特别是，在IP路由器中，令牌桶算法被应用于流量策略及整形。除此以外，使用TCP时为了避免拥塞，所以路由器通常使用一种特殊的反馈机制，叫做随机早期检测（RED）。

20.3.2 令牌桶算法

令牌桶算法是基于分组流与某个预先设定好的平均速率的参考流之比，这个引用流由提供给服务器输入的令牌流代表，它决定何时转发抵达第二个输入口的分组（图20-1）。

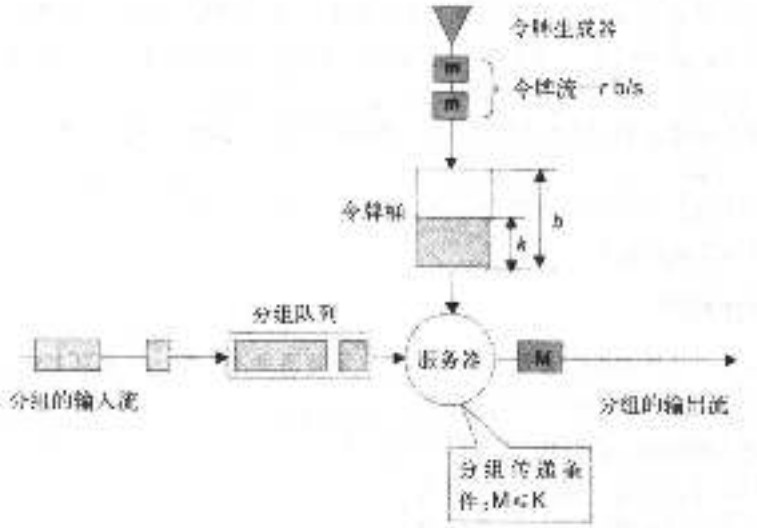


图20-1 令牌桶算法

这种情况下，“令牌”被解释为某些用于构建流量服务模型“端口”信息载体的抽象对象。令牌生成器阶段性地将下一个令牌倒入容量限制为b字节的“桶”中。所有的令牌有着同样大小，等于m字节，并且不停地被产生的这样的桶以每秒r字节的速率填充着。很显然，该速率等于$r = 8m/w$。速率r是被整形流量的最高平均速度，桶的大小对应于分组流量的最大突发大小，如果桶内装满了令牌（即，桶中令牌总数等于b），那么令牌供应暂时停止。当然，令牌桶代表着每w秒增加m的一个计时器。

在流量定义由速率r和突发大小b定义时，何时使用令牌桶算法呢？

参考流与实际流的对比由服务器执行，代表着一种拥有两个输入的抽象设备。输入1连到分组队列，输入2——到令牌桶。服务器也有一个输出，向它传递来自输入队列的分组。服务器的输入1模仿路由器的输入接口，输出则模仿路由器的输出接口。

来自队列的分组只有在抵达服务器时桶内令牌数达到M字节时，它才被服务器转发，其中M为分组大小。

如果满足了这个条件，那么分组被转发到服务器输出端口，并且所有M字节的令牌（精度为m字节）都从桶中移出。如果桶还没有填充到足够的水平，则根据算法使用的目的选择下列两个非标准方法之一来处理分组。

- 如果令牌桶算法被用于整形流量，那么分组只是在队列中延缓一段时间，同时等待足够多的令牌到达桶。因此，可能实现流量平滑：即使突发有大量分组抵达系统，分组依然按照令牌生成器指定的速度均匀地撤离队列。
- 如果令牌桶算法被用于为流量制定轮廓，那么与之相对应的那个分组被丢弃。另外，更加轻松的解决办法是标记分组，降低它进一步服务过程中的状态。比如说，分组可以被标记上一个特殊的删除-合格的属性。此时，一旦发生拥塞，路由器将率先丢弃这个分组。在使用区分服务时，分组被移到另一个低优先级的流量分类并被提供低的QoS。

说明　此算法容忍一定限制范围内的流量突发。假定接口的带宽、速率被令牌桶算法限制为R。然后，对于任意时间间隔t，来自服务器流的平均速率等于下列两个值中的较小值：R以及$(r + b) / t$。如果t值很大，那么输出流速率趋向于r，这就是该算法保证预期平均速率的证据。与此同时，在小段t期间（当$r + bt > R$时），分组会用接口最大速率离开服务器，从而造成突发性流量。此情形发生在分组很长时间未抵达服务器时，导致桶内充满了令牌（即在比b/r更长的期间内）。如果一个大型的背靠背分组序列抵达服务器，这些分组将以输出接口速率R、无间隙地按次序传递到输出口。

这种突发的最大时间为$b / (R - r)$秒，随后需要一个停顿来填满空置的桶。突发大小为$Rb / (R - r)$字节。在已提供的公式的基础上，如果平均速率r选择为近似输出接口带宽的值，可以看出令牌桶算法开始显得难以应付。此时，突发可能会持续相当长的一段时间，从而降低了此算法操作的效率。

20.3.3　随机早期检测

RED算法技术是因特网委员会开发的用以阻止因特网主干拥塞、定义TCP流量的机制。

RED的主要目标是避免严重的网络拥塞。RED与可靠TCP传输协议一起工作并且对于丢失分组使用TCP反作应算法。该反作应包括流量源减缓分组传输入子网的速度。RED使用这个特征作为通知源已经生成太多数据的一个隐性反馈手段。

RED算法使用拥塞层的两个可配置阈值（图20-2）。当拥塞等级降至第一个阈值低阈值（*Lowthreshold*）以下，并不丢弃分组。当拥塞等级降至两个阈值之间，分组丢弃的概率在范围0到可配置值最大丢失概率（*MaxDropProbability*）之间线性增长。在第二个阈值高阈值（*Top*

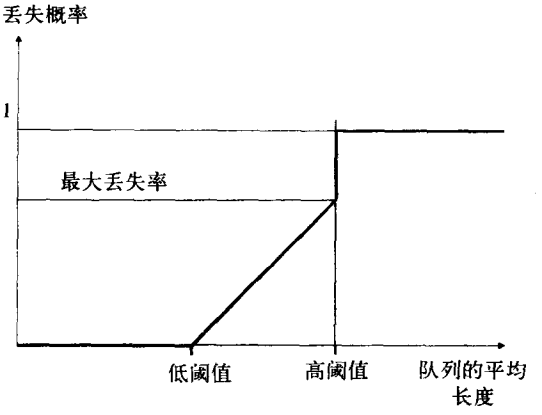


图20-2　使用RED算法丢失分组的概率

Threshold）上正好到达最大丢失概率（图20-2）。当拥塞超过第二个阈值，分组丢失的概率达到100%。

属于特定TCP连接的分组队列的平均长度被用于拥塞度量。

说明 注意对于UDP流量，RED机制并不适用，因为UDP是一个无连接协议，它并不建立逻辑连接而且也不会意识到分组丢失。

当需要为不同的流量分类确保多个反馈参数时，使用加权RED算法（WRED）。RED的这个变形允许为每个流量分类指定独立的低阈值（*LowThreshold*）、高阈值（*TopThreshold*）、丢失概率（*DropProbability*）。通常，WRED机制与WFQ一起使用，确保TCP流量以承诺的速率可靠传输。

20.3.4 集成服务框架和RSVP

集成服务建立的基础是保留沿着端节点到端节点数据流路径上的路由器资源。更加精确地说，在这里所保留的端系统，并不是计算机。这些是运行在端节点上的应用程序（图20-3）。应用程序必须使用适当的API来传递预留特定流资源的请求。这种预留是单向的。因此，如果需要保证双向数据交换的QoS，则需要执行两次预留操作。

集成服务模型中预留利用**预留协议**（reservation protocol，RSVP）来执行。RSVP是一种信令协议，在很多方面类似于**电话网络**（telephone network）中的信令协议。

但是，数据报分组交换网络的特殊特性通常会影响它的操作。因此，IP网络中的参数交换并不代表其预留属性，因为路由器会在路由表（无论有或没有预留）的基础上传递IP分组。

利用RSVP预留所需网络资源，如下面描述的过程来进行，描述中提到的所有报文类型在表20-1中列出：

1）数据源（图20-3中的计算机C1）发送特殊的PATH报文给唯一或组地址（图20-3所示后面一种情况）。它为流量的高质量接收指定了推荐参数：带宽的上阈值和下阈值、延迟、延迟变量。这些流量参数包含在流量规范（*traffic specification*，*TSpec*）中。PATH报文根据任意路由协议所获得的路由表（如，OSPF）被网络路由器送往目的地（或多个目的地）。为了指定流量参数，使用令牌桶算法的参数（如，平均速率和桶深度）。除此以外，流的最大允许速度以及分组大小限制可以额外地指定。

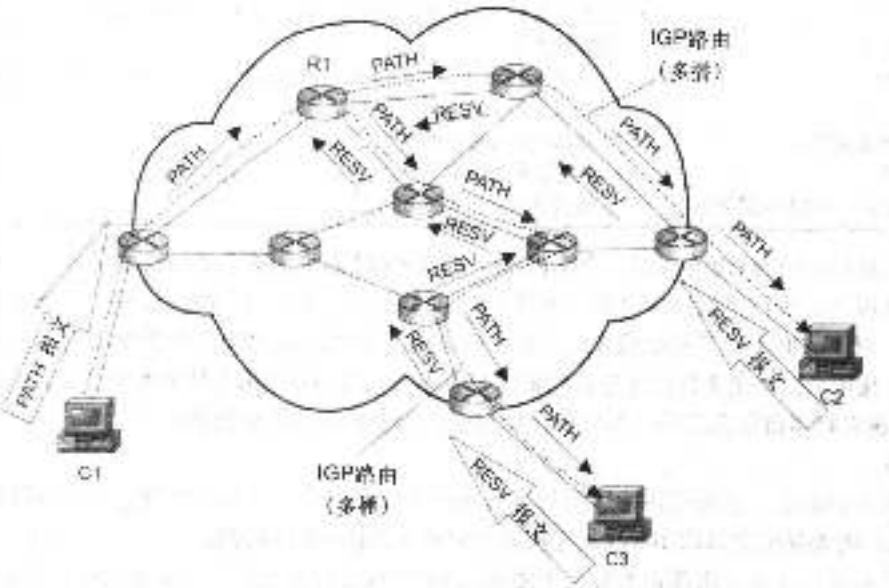


图20-3 使用RSVP的资源预留

2）每个支持RSVP的路由器收到PATH报文之后，记录“路径状态”，其中包含了PATH报文源以前的地址——反方向中的下一步（即，朝着源的方向）。需要确保接收端的响应遵从与PATH报文同样的路径。

3）收到PATH报文之后，接收端（目的节点上运行的应用程序）发送RESV报文资源预留请求给发送给它此消息的路由器。RESV报文请求保留资源。图20-3显示了两个接收端，计算机C2和C3。除了TSpec报文以外，RESV报文还包含请求规范（*request specification*，*RSpec*），其中指定了接收端所需的QoS参数、以及过滤说明（*filtering specification*，*filterspec*），它定义了此保留应该应用于连接的哪些分组（比如，通过通信协议类型和端口号）。RSpec和filterspec的组合叫做流描述符（*flow descriptor*），路由器用它来识别每个资源保留。RSpec说明书中的请求QoS参数可以与TSpec中的不同。比如说，如果接收端决定不要接收源发送的所有分组而是选择性地接收其中一部分（可以使用过滤说明来指定），那么所需的带宽就会少一点。

4）当沿着路由支持RSVP的任意路由器接收到RESV报文，向上行流的方向传输时，它使用两个过程来决定是否接收请求中指定的预留参数。利用授权控制机制，它检查路由器是否拥有支持所申请QoS水平的资源。策略控制过程被用于检查用户是否拥有资源保留的权限。如果因为缺乏资源或者权限错误而导致不能满足请求时，路由器返回一个错误消息给发送端。如果请求被接受，那么路由器沿着到下一个路由器的路由发送RESV报文，并且有关被申请QoS水平的数据被传递到负责流量控制的路由器机制。

5）路由器接收预留申请还意味着QoS参数被送往路由器的合适单元以供处理。QoS参数处理的特定方式在RSVP中并没有描述，但是，通常它包含路由器上空闲带宽以及新预留所需内存的可用性。如果检查结果为肯定的，那么路由器存储新的预留参数并且将它们从合适源对应的计时器值中减去。

表20-1　RSVP报文

报文类型	内　容
源到目的的PATH报文	源的流量说明
源的流量说明	高质量流量接收的推荐参数：高和低带宽限制，延迟和延迟参数，令牌桶算法参数——平均速率和桶深度。另外，可以申明最大允许速率和流分组的阈值大小
过滤说明	申明此预留必须被应用于会话的哪个分组（比如，根据传输协议类型以及端口号）
接收端请求说明	接收端需要的QoS参数
流描述符	流量说明 + 接收端请求说明
RESV报文——请求保留资源	源流量说明 + 流描述符

6）当沿着上行路由的最后一个路由器收到RESV报文并且接受了该请求时，它向源节点发送一条确认报文。执行了组预留之后，传送树的分支点被汇合成一个加以考虑。打个比方，考虑的例子中，RESV的C1和C2接收端被汇合到路由器R1。如果为所有的预留流申请同样的带宽，那么也为汇合流申请。如果为预留流申请的带宽大小不一，则为公共流选择最大的带宽。

7）建立起预留状态之后，源开始传输数据；沿着接收端的全程路由，以指定的QoS标准提供数据服务。

需要强调的是，上面描述的过程只执行单向的预留。为了在用户连接架构中确保数据传输的QoS参数，需要保证发送端和接收端互换角色并且在此执行RSVP操作。

为了对数据流量应用预留参数，需要确保RSVP报文以及数据分组沿着同样的路由通过网络。如果RSVP报文根据用户流量路由表的相同记录的基础上传输时，就可以保证这一点。

说明 如果RSVP报文以从路由表中选择恰当记录的传统方法传输，因为所有可能的路由不用于预留，那么就失去了流量工程会话选择的可能性。相反，只有根据一些路由协议的路由度量值选择的最短路径会被用于此目的。

可以直接或间接取消预留。直接取消由发送端或接收端利用RSVP的适当报文来执行。间接取消则在超时后发生：预留状态有一个特定的TTL，正如路由表中的动态记录那样。根据RSVP，接收端必须阶段性地确认预留。如果不再收到确认报文，那么当它的TTL过期时就取消预留。这种取消叫做软取消。

RSVP开发了很多拓展功能，这些拓展使得该协议超出RSVP架构范围内的操作。其中最重要的解决方案是流量工程拓展，用于MPLS中并且概括了第五部分中提到的MPLS技术。

20.3.5 区分服务框架

区分服务与集成服务基于相同的QoS概念，然而，区分服务将流量分类视为服务对象而不是单独的流。

流量分类（traffic class）是被有相同标签的一个分组集——举个例子，这些可能是语音应用的所有分组或者通常最小MTU的所有分组。

与数据流相比，流量分类并不根据路由识别分组。

图20.4说明了这个区别。因此，路由器R1将所有需要优先级服务的流和到达接口i1的流归为相同的流量分类，不论它们后续的路由是什么。路由器R2操作另一个优先级分类，因为它并不包含路由器R1的接口i1的所有流。

通常区分服务网络支持少量流量分类的区分服务——举个例子，两个（延迟-敏感的和弹性的）或三个（除了延迟-敏感和弹性流量外还支持另外一个流量分类，它需要保证以预设的最小流量速度传送分组）。少量流量分类确保了这个模型的可延拓性，因为路由器并不需要存储每个单独用户流量的状态。而且，每个路由器可以决定提供给特定聚合流的服务，这样保证了区分服务的高可延拓性。此决定独立完成并且不需要与其他路由器协调，这种解决方案叫做*单跳行为*（*per hop behavior，PHB*）。

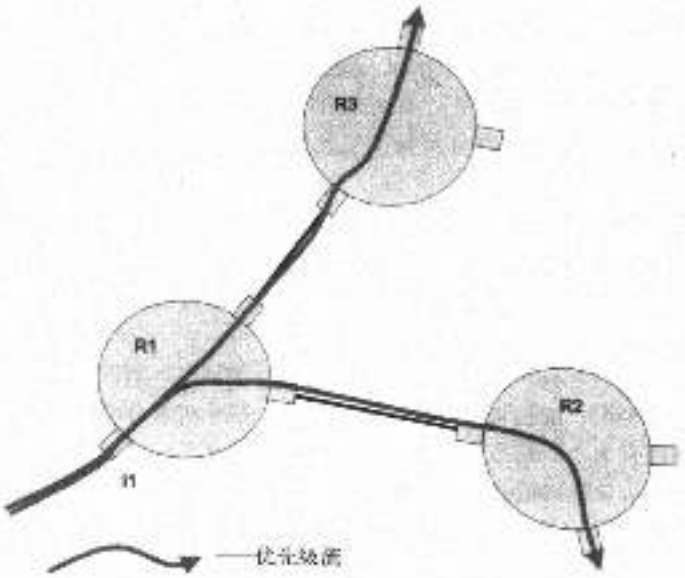


图20.4 与集成服务（IntServ）相比，区分服务（DiffServ）将聚合流而不是单个流视为服务对象

区比，既然区分服务架构并不跟踪分组路由，所以它不能使用资源预留协议。这点与综合服务架构中的RSVP相似，取而代之，网络路由器为网络支持的每个流量分类执行静态资源预留。

在区分服务中，*IP*优先级（*IP Precedence*）字段或它的后继字段携带的标签，区分服务字节（*DiffServ byte*）字段，被用于指定属于特定流量分类的IP分组。

正如图20-5中所示，虽然*DS*字段占用了*TOS*字段，但它在有关RFC（791、1122、以及1349）中重新定义了该字段的位值。当前，只使用了*DS-byte*字段的高六位，其中最高二位用于定义流量分类（提供了不超过8种不同的分类）。最低位（在使用的6位中）通常携带*IN*属性，显示分组"脱离"了流量定义（类似于被延迟技术的*DE*属性或ATM技术的*CPL*）。中间两位描述了相同流量分类范围内分组服务的不同种类。

支持区分服务的路由器必须支持流量分类、标记、度量、条件，以及流量平滑和优先队列或加权队列中的服务。

虽然每个网络路由器都可以标记分组，但区分服务模型认为在网络入口点标记分组是十分流行的。此网络入口点必须支持区分服务协议并且处于单个组织的管理控制下。这种网络叫做区分服务域，在分组离开区分服务域时，移除标记这样，另外一个域可以重新标记。区分服务边界路由器扮演着域检查点的角色，它们检查输入流并且判断它是否具备区分服务的能力。

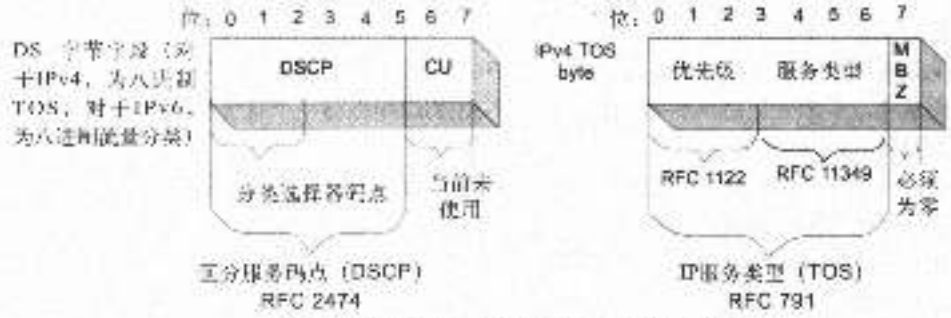


图20-5　DS字节与TOS字节字段的对应关系

区分服务协议假定有着公共边界的域之间存在着服务许可协议（SLA）。SLA指定策略标准并给出了流量轮廓。通常期待流量会根据SLA在域输出点形成并整形，而在域入口点流量则根据策略规则整形。定义范围外的任意流量（比如，超出SLA指定带宽最大限制的流量）都不能获得服务保障（或简单地根据SLA支付高速率）。策略标准可能包括一天中的时间、源和目的地址、通信协议以及端口号。当审视策略规则并且流量满足预设的定义时，区分服务域必须确保SLA中定义的流量的QoS服务。

如今，IETF已经开发了两个逐步分组转发（PHB）的标准，它们代表着两种不同的服务：

- *加速转发*（*Expedited forwarding*，*EF*）。这种服务类型的特征为单个代码值（10111）。它提供了最高QoS等级、最小化的延迟和延迟变量。任何超过流量定义中指定强度的流量都将被丢弃。
- *确保转发*（*Assured forwarding*，*AF*）。在这种服务类型中，有四种流量分类，每个分类中有三个等级的丢弃分组，加起来总共12种流量类型。每个流量分类被分配一个预设的最小带宽和存储队列的缓存大小。超出定义的流量以比满足定义条件更低的可能性传递。这意味着可以为超出定义的流量提供低质量服务而没有必要丢弃掉。

基于这些逐步说明以及恰当的SLA，可以向端用户提供端对端服务——分别是EF服务和AF服务。

EF服务的主要目标是提供与专用线路的QoS相匹配的QoS服务。由于这个原因，这种服务又叫做"虚拟专用线路服务"，同时又被称为在区分服务IP网络中强调最高QoS的"高端服务"。随后，

这个说明被RFC 3246废弃。它给出了更加准确的EF服务定义。

如果EF服务的实施机制允许无限制地丢弃其他流量（比如，优先级队列），那么它的实现必须包括一些工具，以限制其他流量分类上的EF流量所造成的影响。比如说，这可以通过利用令牌桶算法来设置限制路由器输入的EF流量速率来实现。EF流量的最大速率以及可能的突发大小必须由网络管理员设置。

AF服务的四个组面向确保传送的顺利进行，然而，并不具备EF服务约定的将分组延迟等级最小化。传送只在输入流量的速率不超过为该分类分配的最小传输带宽的情况下才得以保证。AH服务的实现可以与EF服务很好地结合，因为EF流量可以根据优先级方法使用，只是输入流的强度有限。剩余带宽根据加权队列处理算法分配给AF服务的流量分类，这将确保所需的带宽。然而，并不会最小化延迟。AF服务的实现意味着（但不是要求）为每个分类使用加权队列处理并使用RED反馈。

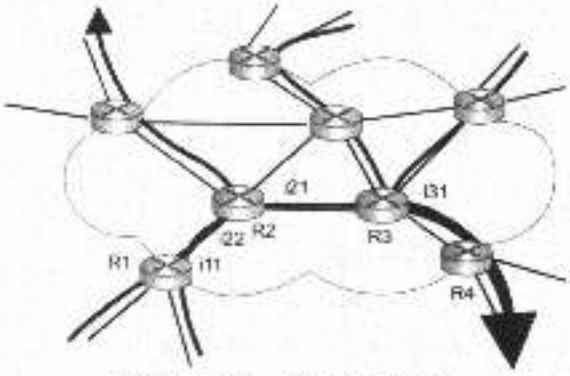


图20-6 DiffServ服务的不确定性

利用区分服务进行流量服务的相对简单性决定了它的缺陷。主要缺陷是很难向服务使用者提供质量保证。请结合图20-6中显示的网络考虑这个缺陷。

服务流量分类假定边界路由器监督流量时并不考虑分组的目的地址，这个解决方案可以被称做目标不警觉。通常，对于边界路由器的输入接口，为每个流量分类指定所允许工作负载的预设限制。举个例子，假设网络服务于两类流量，高端（*premium*）和尽力服务（*best effort*）。高端流量的阈值为20%，每个边界路由器的所有输入接口都设置为这个值。除此之外，假设为了简单起见，所有网络路由器的所有接口都有着同样的带宽。

很明显，虽然有这个相当苛刻的限制，但网络路由器的接口还是有着不同的负载。为了简单起见，图20-6只显示了需要高端服务的流量。因此，路由器R1的输出接口i11服务于两个高端服务流，负载为40%，而路由器R2的输出接口i21只服务于一个这样的流，因为第二个流从另一个输出接口传输。至于路由器R3的输出接口，则超负载了，因为它服务于三个高端服务流，因此，利用率为60%。在这些影响队列产生因素的基础上（参见第7章），得知利用率是最重要的因素，关键值为大约50%。因此，高端服务分组的长队列将产生在接口i31，这些队列会降低QoS，因为它们导致相当多的延迟甚至导致分组丢失。尽力服务分类的流量也将被影响，因为它仅可使用40%的接口带宽。

很自然，我们简化了这个模式，因为主干路由器的接口通常比边界路由器的更快，因此，它们的使用率将比输入接口使用率的总合更低，正如例子中的情况。为了减少主干路由器的内部接口以及边界路由器的输出接口上超载的可能性，也可以降低输入接口上允许的高端流量负载，比如，降低5%。

然而，这些措施并不能保证所有网络路由器的所有接口都以所需范围的利用率操作，从而确保所需的QoS。为了提供这样的保证，需要利用流量工程方法——举个例子，控制流量分流而不是分类，或者在这种情况下，聚合流。聚合流（aggregate flow）是包含相同分类分组并拥有穿越子网相同部分路由的流。这个公共部分并不包括从边界路由器输入接口到另一个边界路由器输出接口的完整路径。对于分组，经过至少两个公共接口就足以被认为是聚合流了。例如，这就是图20-6中经过接口i11和i22的流。

然后，已知每个聚合流穿过网络的路由，就可以检查路由沿途是否有足够的资源供每个流使用。比如说，需要检查是否接口利用率超过了预设的阈值。为了实现这点，需要使用分组目的地址计数的策略——如目标-警醒策略（*destination-aware policing*）。但是，在IP网中实施这种解决方法将面临几个困难。首先，区分服务技术并不使用诸如集成服务技术那样的RSVP信令协议，这意味着路由器上对每个聚合流进行的资源可用性检查必须在脱机模式下手动地或利用一些特殊软件执行。其次，为了执行这种计算，需要知道穿越网络的流路径。这类路由是根据某路由协议构建的路由表决定的，比如RIP或OSPF；当网络中使用IGP分类的多个路由协议时，由路由协议的联合决定；或者，手动配置。因此，对于手动或自动计算，需要知道所有网络路由器的路由表并且跟踪它们的变化。这不是一个微不足道的任务，考虑到网络链路可能发生故障或路由器重建这类表。而且，需要记住路由器可以应用负载均衡方法，将聚合流分割成多个流，这同时也使得计算更加复杂。

区分服务的目标-警醒版本提高了通信载体提供的QoS，但是，它也使此方法的理念复杂化，因为该方法建立在每个网络路由器独立服务流量分类的基础上。

20.4　网络地址转换

互联网中的路由基于分组头部的目的地址执行。通常，这些地址自从被发送端形成就保持不变，直到到达目的节点。然而，这个原则也有例外。举个例子，在广泛使用的**网络地址转换**（network address translation，NAT）技术中，通常假设根据公司内部网中分组路由所使用的地址，在外部因特网中转发分组。

20.4.1　地址转换的原因

使用NAT技术最常见的原因是IP地址短缺。如果出于某原因需要联入因特网的公司不能从提供商得到所需数目的全局IP地址，它就可以使用NAT技术。这种情况下，**保留（私有）地址**（reserved (privato) address）被用于定为内部地址。我们曾在第17章描述过它们。

为了使具有私有地址的主机能够使用因特网通信或是与拥有全局地址的主机进行连接，就需要用到NAT技术。

当一个公司为了安全考虑需要隐藏内部网中的主机地址时，就证明了NAT技术是有用的。它阻止了入侵者获知网络结构和规模以及输入输出流量的强度。

20.4.2　传统的NAT

NAT技术有好几种变化形式，其中最流行的是**传统NAT**（traditional NAT），它允许私有网络的主机以对用户透明的方式去访问外部网络的主机。需要强调的是，NAT的这种形式解决了只能组织**输出**（outgoing）连接会话的问题，此时的会话方向根据其发起者的位置来决定。如果数据交换由运行在内部网络中的某主机发起，那么该会话被称为输出，即使该会话过程中的外部数据可以被传送进内部网络[⊖]。

NAT的思想基于下列的解决方案（图20-7）：假设公司网络形成了一个桩域，其中的主机都被指派了私有地址。NAT软件安装在将公司网络连接到外部网的路由器上。NAT将私有IP地址集{IP*}自动映射到全局IP地址{IP}，全局IP地址由公司从ISP获取并指派给公司路由器的外部接口。

NAT操作的一个重要特征为如下规则，即路由器广告沿着私有网络边界传播。外部网络的路

⊖ 传统NAT只在一种例外情况下才允许反方向会话的存在。为实现此目的，它利用到某有限主机集合的映射，其指定了内部地址到外部地址唯一且确定的对应关系。

由协议广告被边界路由器传递到内部网络，交给内部路由器处理。然而，相反的陈述并不符合事实。外部网络路由器无法收到内部网的广告，因为这种广告在传递信息给外部接口时被过滤掉了。对此，内部路由器“了解”所有外部网络的路由，而外部路由器对私有网络却一无所知。

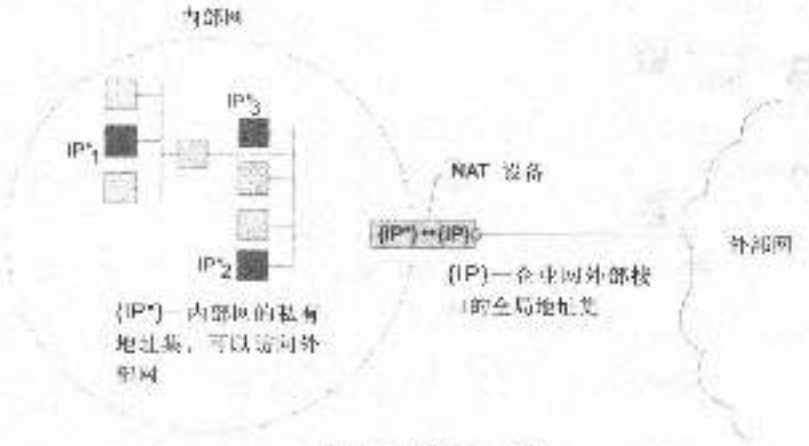


图20-7 传统NAT方法

传统NAT被划分为**基本网络地址转换**（Basic network address translation，基本NAT），此方法指使用IP地址实映射，以及**网络地址端口转换**（network address port translation，NAPT），此方法使用地址映射和传输标识符。最常见的情况下，TCP/UDP端口被用做传输标识符。

20.4.3 基本的NAT

如果负责访问外部网的本地主机数不超过可用的全局地址数，那么就可以保证私有地址和全局地址之间的唯一映射。在任何时候，都可能发生与外部网络交涉的内部主机数被可用全局地址数所限制。此时，使用NAT的主要目的是确保安全性而不是解决地址短缺的问题。

一些主机的私有地址可以静态地映射到全球地址。这类主机可以使用分配给它们的全球地址从外部访问。内部和外部地址的映射由安装了NAT软件的网络路由器或任何其他设备（比如防火墙）支持的路由表指定。

封闭域可以使用私有地址匹配。比如，网络*A*和*B*（图20-8）使用同一块地址进行内部寻址：10.0.1.0/24。与此同时，两个网络的外部接口地址（网络*A*中的181.230.25.1/24、181.230.25.2/24以及181.230.25.3/24和网络*B*中的185.127.125.2/24、185.127.125.3/24、185.127.125.4/24）是全局唯一的。因特网上没有任何其他主机使用它们。在此例中，每个网络中只有三个主机可以突破公司网络的限制。这些主机的私有地址到全局地址的静态对应关系在两个网络边界设备上的路由表中指定。

当网络*A*的主机10.0.1.4发送一个分组给网络*B*的主机30.0.1.2，它在分组头部的目的地址字段中放置全局地址185.127.125.3/24。源节点发送分组给它的默认路由器*R1*，它知道到网络185.127.125.0/24的路由。路由器将分组传递给边界路由器*R2*，后者也知道到网络185.127.125.0/24的路由。在发送分组之前，此边界路由器上运行的NAT协议利用映射表将源地址字段中指定的私有地址10.0.1.4替换成对应的全局地址：181.230.25.1/24。当这个分组沿着因特网传送到达网络*B*的NAT设备的外部接口时，全局目的地址185.127.125.3/24被转换成私有地址10.0.1.2。反向传送的分组进行地址转换的相同步骤。

注意，前面描述的操作中并不需要发送端和接收端节点的参与，这意味着整个过程对于用户来说是透明的。

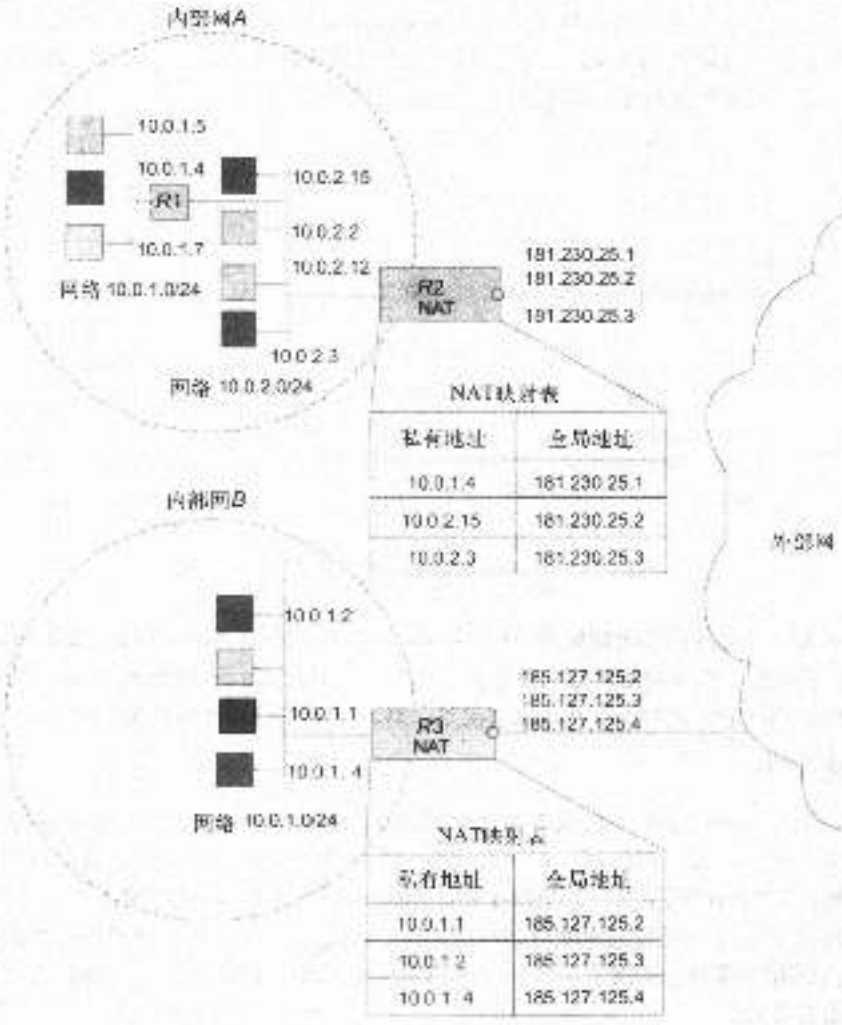


图20-8　基本NAT：输出会话的地址转换

20.4.4　地址和端口转换

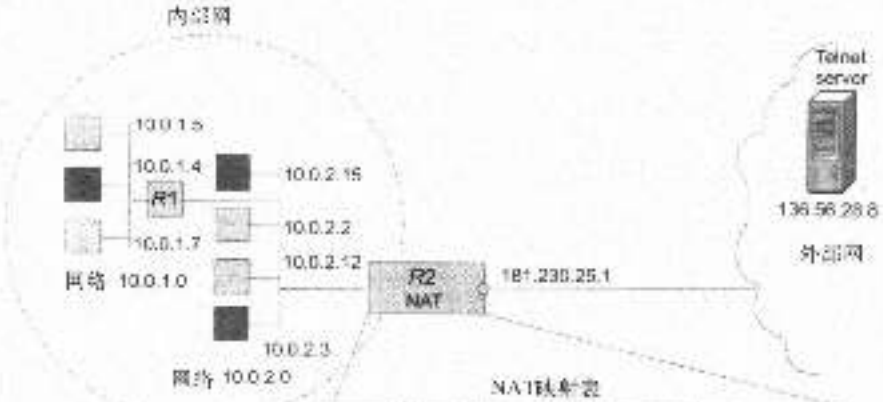
假设某些组织拥有通过WAN链路连接到ISP网络的私有IP网。为边界路由器*R2*的外部接口指派了一个全局地址，并且企业网的所有其他主机都被指派了私有地址。NAPT允许内部网的所有主机利用单个注册IP地址同时与外部网通信。这里就会问一个很自然的问题：用来响应此私有网络所发送请求的外部分组如何找到发送端主机？毕竟，发送到外部网络所有分组的源地址字段都包含着同样的地址，也就是，边界路由器外部接口的地址。

为了唯一标识发送端主机，使用了额外信息。如果IP分组封装了UDP或TCP的数据，那么UDP或TCP端口号就被用做这种额外信息。然而，即使这样也不能完全澄清事实，因为内部网络可能发起多个匹配发送端端口号的请求。因此，又遇到了如何确保单个全局地址唯一且确定地匹配到内部私有地址的问题。解决方案在于当分组从内部网传送到外部网时，每对{*内部私有地址；发送端TCP/UDP端口号*}（{*internal private address; sender TCP/UDP port number*}）被映射到对{*外部接口的全局IP地址；指定的TCP/UDP端口号*}（{*global IP address of the external interface; assigned TCP/UDP port number*}）。指定的端口号可以任意选择，然而，它必须是有权限访问外部网络主机

中的唯一的一个。这种映射注册在表中。

此模型满足了大多数访问外部网络的中型网络的要求，它们利用从提供商获得的单个注册IP地址来实现。

图20-9显示了一个示例，其中封闭网络A使用地址块10.0.0.0的内部地址，此网络路由器的外部接口地址为181.230.25.1，由ISP分配。



NAT映射表

私有地址	端口	全局地址	已分配端口
10.0.1.4	1245	181.230.25.1	3451
10.0.2.15	1245	181.230.25.1	3452
10.0.2.3	1045	181.230.25.1	3455

图20-9 NAPT：输出TCP/UDP会话的地址和端口号转换

当内部网的主机10.0.1.4发送分组给外部网中的telnet服务器时，它使用全局地址136.56.28.8作为目的地址。分组抵达了路由器*R1*，此路由器知道到网络136.56.0.0/16的路由经过边界路由器*R2*。路由器*R2*的NAPT将地址10.0.1.4和源TCP端口1245转换成全局唯一地址181.230.25.1和唯一分配的TCP端口（此例中为3451）。分组以这种形式被传送到外部网络并到达telnet服务器。在接收端生成响应消息之后，它将内部网中唯一注册的全局地址指定为目的地址，这就是NAPT设备外部接口的地址。至于接收端端口号，服务器从收到分组的发送端端口号字段中提取出已分配的TCP端口号，指定给它。当确认分组到达内部网的NAPT设备时，利用端口号从转换表中选择所需的行。此行定义了适当主机的内部IP地址和它实际的端口号。此转移过程对于终端节点来说完全是透明的。

说明 注意转换表包含端口号为1245的另一条记录。这种情形也是可能的，因为运行在不同计算机上的操作系统独立为客户端程序指派端口号。正是这种不确定消除了唯一分配端口号就重用的可能性。

NAPT的变化形式只允许由私有网络发起的TCP/UDP会话。然而，也会存在需要从外界访问内部网某节点的可能。最简单的情况下，当服务已经被注册——即，服务已经分配给它一个知名端口号（如，Web或DNS）——并且此服务在内部网中表示成一种简单的形式，那么这个问题相对来说容易解决。此服务以及操作它的主机根据服务的已注册的知名端口号唯一地定义。

总结NAT技术所涉及到的内容，可以注意到除了传统NAT，还有一些NAT的变化形式。其中一个示例就是双重NAT，源和目的地址都发生了改变，只有一个地址发生改变的传统NAT相比。在私有和外部地址空间发生冲突时，双重NAT是必须的。大多数情况下，这发生在当内部域拥有属于另外一个组织错误分配的公共地址时。这种情形也会发生在组织网络最初与外界隔离开时，

此时地址——从全局地址空间中提取出的——是任意分配的。当组织想为内部网主机保存旧地址时，有时会因为ISP的改变而引起这种冲突。

20.5 路由器

20.5.1 路由器功能

路由器的主要功能是读取收到网络分组的头部并且缓存在每个端口上（比如，IPX、IP、AppleTalk或DECnet）。此后，根据分组的网络地址，路由器选择它的路由。因此，此地址包括网络号和主机号。

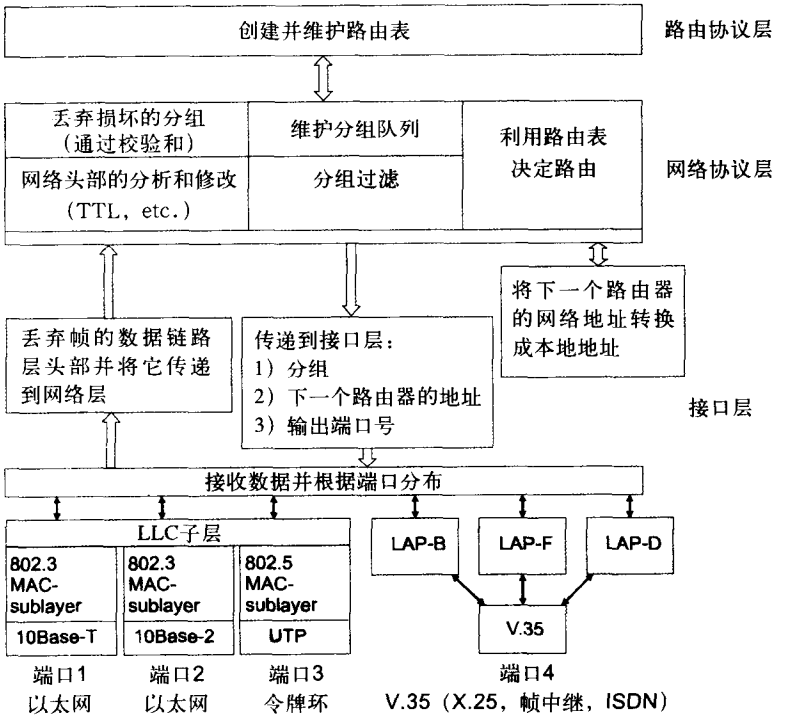


图20-10　路由器的功能模型

路由器功能可以根据OSI模型层次分为三个组（图20-10）。

1. 接口级

在较低层上的路由器，如同任何连接到网络的设备一样，确保连入传输介质的物理接口，包括协调电路信号等级、线路和逻辑编码、为路由器配备特定类型的连接器。在路由器的不同模型中，通常提供的不同物理接口集合是连接LAN和WAN的各种组合端口。特定数据链路层协议与其接口不可分割地联系在一起——比如，以太网、令牌环或FDDI。连接到WAN的接口通常只决定某些物理层标准。基于这些标准，多个数据链路层协议可以在一个路由器中操作。举个例子，WAN端口可以支持V.35接口，基于它多个数据链路层协议可以操作：PPP（传输IP流量和其他网络层协议）、LAP-B（用于X.25网络）、LAP-F（用于帧中继网络）、LAP-D（用于ISDN网络）以及ATM。LAN技术可以解释LAN和WAN接口之间的区别，它决定物理层和数据链路层协议只有组合起来才能使用。

路由器接口执行整套与帧传输相关的物理层和数据链路层功能，包括访问媒质（如果需要的话）、形成位信号、接收帧、计算校验和以及如果校验和的值正确时传递帧数据字段给上层协议。

说明 如同任何普通的终端节点一样，每个路由器端口都有它自己的硬件地址（在LAN里，这是MAC地址），通过此地址所有其他的节点发送需要路由的帧。

特定路由器模型支持的物理层接口列表对于消费者来说是最重要的特征。路由器必须支持所有它直接连接每个网络的数据链路和物理层协议。图20-10显示了一个路由器的功能模型，该路由器的四个端口实施了下列物理接口：两个以太网端口10Base-T和10Base-2，令牌环的UDP，LAP-B、LAP-D、LAP-F协议可以操作的V.35，确保了连接到X.25、ISDN或帧中继网络的可能性。

提供给路由器端口的帧，被恰当的物理和数据链路层协议处理后，从数据链路层头部脱离。从帧数据字段检索的数据被传递到网络层协议实体。

2. 网络层协议

与之相反，网络层协议从网络层分组头部检索并且**分析更正该字段的内容**（analyzes and corrects the contents of its field）。首先，需要测试校验和；如果分组受损，必须将之丢弃。然后再检查分组花费在网络中的时间是否超出了允许的阈值（TTL）。如果超出了TTL，也将分组丢弃。在这个步骤中，所有必需的更正信息被输入到字段的内容中——比如，需要的话，增加分组的TTL或者重新计算它的校验和。

最重要的路由器功能，**过滤**（filtering），也由路由器的网络层执行。为网桥或路由器而封装在帧数据字段中的网络层分组由一串无结构的二进制序列代表。另一方面，路由器软件包含网络层协议实体，因此能够分析帧并且**解析独立的字段**（analyzing individual field）。此软件配备高级GUI工具，使得管理员可以正确地制定复杂的过滤规则。通常，路由器也能够分析传输层报文的结构。因此，过滤器会阻止某应用程序服务的报文，比如telnet服务，进入网络中。这种过滤通过分析传输报文中的协议类型字段来执行。

路由器的主要路由器功能，名义上，为决定路由，也与网络层相关。网络层协议实体利用从分组头部检索出的网络号来找到路由表中的某一行，其中包含了下一个路由器的网络地址以及需要传递分组以确保它朝正确方向传递的端口号。

在下一个路由器的网络地址被传递到数据链路层之前，它必须根据包含的下一个路由器的网络中所使用的网络技术转化成本地地址。为了实现此目的，网络协议申请地址解析协议（*address resolution protocol*，*ARP*）。

从网络层，下一个路由器的本地地址和路由器的端口号被传递到数据链路层。基于特定的端口号，分组被传给其中一个路由器接口，分组然后以恰当形式的帧封装起来，下一个路由器的本地地址被放置在帧头部的目的地址字段中。随后，将帧发送给网络。

3. 路由协议层

网络协议在操作过程中积极地使用了路由表。然而，这些并不包含创建或者维护它的内容。这些功能由路由协议完成，基于这些协议，路由器交换网络拓扑的信息，然后通过分析收到的数据来判断满足特定标准的最佳路由。分析的结果构成了路由表的内容。

除了前面列出的功能外，也可以为路由器指派其他功能——比如，与分片相关的操作。

20.5.2 路由器按应用范围的分类

根据应用范围，路由器划分为多个分类（图20-11）。

主干路由器（*backbone router*）旨在构建通信载体或大型企业的核心网络。主干路由器操作携带多个用户连接数据的聚合信息流。

主干路由器的主要目的在于创建网络的高性能且可靠的交换核心。为了执行这个任务，主干路由器配备了高性能接口，比如ATM 155/622 Mb/s、千兆位以太网、10千兆位以太网以及

SONET/SDH接口以确保从155Mb/s到10Gb/s的速率。为了创建容错核心拓扑，主干路由器必须支持多个这样的接口。

很自然，为了避免造成网络核心的瓶颈，主干路由器必须提供非常高的性能。举个例子，如果路由器配备了8个10Gb/s接口（以太网或SDH），那么它总共必须能承担80Gb/s的性能。为了达到这样的性能，主干路由器仿照LAN交换机的结构分布其内部架构，这在第15章中已经介绍过。基于路由表的本地备份，每个端口或端口组配备自己的处理器来传递IP分组。为了在端口之间传递分组，使用基于共享内存、公共总线、电路交换的交换结构。与LAN交换机相似，高性能的路由器有一个集中处理单元执行一般的任务：构建路由表、存储配置参数、支持路由器的远程管理等等。因为与传递IP分组相关的功能比LAN技术（如以太网）中的帧传递复杂得多，为路由器开发高性能端口的任务比交换机的类似任务难度也大很多。为了使端口处理器简单并便宜一点，开发者通常并不用它执行附加的路由器功能，比如流量过滤或NAT，即使QoS支持也不完全由这种处理器实现。通常，只实现队列机制，不实现流量定义。这是因为主干路由器只在网内操作，不与外部交涉，这意味着它并不执行需要定义和过滤的边界功能。这种路由器的主要任务是以最大速度在接口之间传递分组。

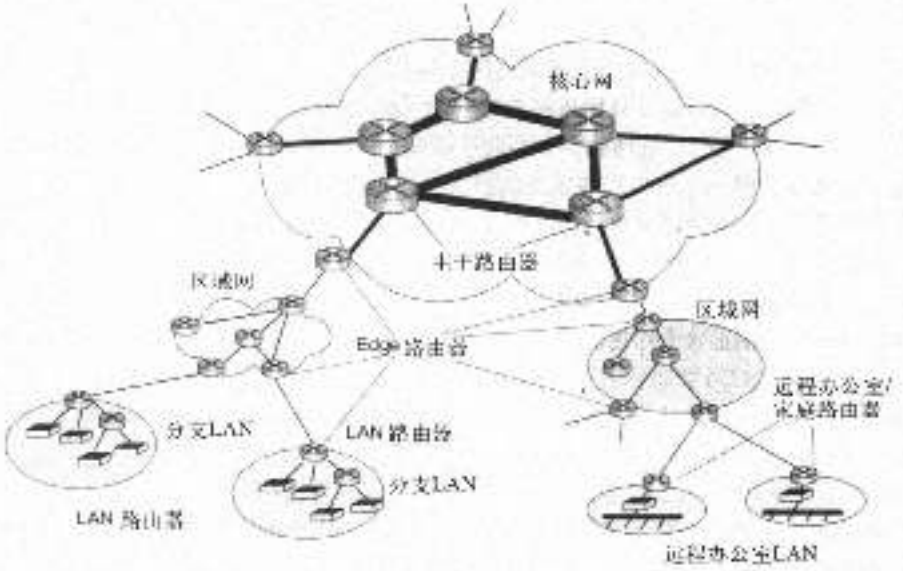


图20-11　路由器类型

大量接口允许构建与全连接设计相近的冗余拓扑，以此来保证网络容错性。但是，边界路由器必须自己提供高可靠性。路由器的可靠性和容错性是付出使用冗余模块如中央处理单元、端口处理器、电源供应单元的代价才实现的。

边缘路由器（*edge router*）连接主干和周围网络，这些路由器形成执行从与主干相关的外部网中接收流量的功能。边缘路由器又叫做访问路由器。周围网络通常是自治的，可能是直接连到主干或拥有自己主干的大型企业或不同部门网络的通信载体客户端。

无论哪种情况，来自主干的管理员无法控制的网络的流量都可以抵达边界路由器的所有接口。因此，必须过滤掉并且监督这种流量。这导致边缘路由器的需求与主干路由器的需求不一样。路由器通过实现额外流量过滤以及监督功能来确保最大灵活性的能力被提到了案前。而且，确保边界路由器的性能不能因为这些额外的功能受到影响，这一点非常重要。边界路由器接口不如主干路由器的快，但是，边界路由器更加灵活，因为它是基于不同的技术连接到主干网络的。

将路由器划分为主干和边缘并不是严格和确定的。这种划分仅仅反应了路由器更倾向于使用以及最能体现优势的应用领域。任何路由器都可以应用在它的主应用领域范围之外。举个例子，配备了高性能端口的主干路由器可以同时扮演边缘路由器的角色。执行大型网络边界路由器的任务，很好地胜任此角色的路由器也可以为小型网络的主干路由器服务，这里它的接口能够应付这种网络上主干的负载。

将路由器划分为主干和边缘路由器仅仅反应了它们应用的一个方面，它们在本地和外部网中的位置。很显然，还存在着其他方面。举个例子，可以将路由器划分为在通信载体的路由器和公司网络中使用的路由器。

载体路由器（carrier router）与其他类型路由器的主要区别在于高可靠性和作为因特网组成部分的商业操作所需的完整函数集的支持，从BGP到控制用户数据流的系统，这些都是计费系统所需要的。在提供商业服务时，路由器失效所付出的惨重代价说明了可靠性的重要，对于数据传输服务可靠性的需求持续增长；因特网和VPN用户希望这些服务与电话通信一样可靠。因此，当我们说路由器的可用性达到0.999的边界，非常接近于电话设备的可靠性参数0.9999时，这对于主干和边缘来说，都是主要与通信载体路由器相关的。

企业路由器（corporate router）旨在公司网络范围内使用；因此，对可靠性的需求比通信载体路由器的要低。而且，企业路由器不需要作为独立、自治系统操作所需的完整功能集。

很自然，通信载体和企业路由器的特征很大程度上依赖于通信载体或企业的规模以及它们特定的特征。如今，一个与ISP层次中Tier1相关的国际通信载体需要配备10Gb/s接口的主干路由器。这种路由器将很快被配备了DWDM端口的路由器取代，后者用40波操作并确保总共400Gb/s端口的通信速率。这种载体的边缘路由器同时也具备顶端模型的资格，此类的路由器拥有确保访问速率为622Mb/s——2.5Gb/s的端口。

小一点的通信载体，比如区间的和本地的，并不需要这种性能水准的路由器，因为它们的流量总合相对小很多。因此，这种载体的主干路由器可以限制为支持2Mb/s——155Mb/s接口，而且边缘路由器必须确保可以利用电话线进行拨号访问。在小型网络中，可能没有主干路由器，因为这种网络通常包含一个或多个边缘路由器。

类似情形存在于企业网中，这里可以找到不同可靠性和性能等级的路由器。举个例子，大型企业可以使用主干和边缘路由器，特征与Tier1通信载体的路由器相近。但是，企业网通常基于低一层特征的设备而建。这意味着大型企业使用与区间通信载体相似的设备，以此类推。

部门路由器（department router）负责区域性部门之间以及与核心网络的连接。区域部门的网络，与核心网络相似，可以包含多个LAN。这种路由器通常是某一企业主干路由器的类似版本。

如果部门路由器建在机架基础上，那么此机架中槽数为四或五。也可能是固定端口数的设计。支持的LAN和WAN接口速度比企业路由器的慢，已公布路由器中这个分类是最大的。它们的特征可以在从与主干路由器相近到远程办公室路由器的特征范围内变化。

远程办公室路由器（remote office router）只负责利用WAN链路实现远程办公室的LAN到核心网络或到区域部门网络的连接。

通常，LAN接口为以太网10/100Mb/s，并且WAN接口是确保速度为64Kb/s，1 544Mb/s或2Mb/s的租用线路。远程办公室路由器能够支持利用拨号电话线作为租用线路保留链路的操作，有很多种远程办公室路由器。这是因为潜在客户数量很大而且存在大量这种设备的不同应用程序。大量潜在客户的不同特征以及这种设备的详细说明可以解释这点，其中一些为支持长距离通信的特定类型。例如，存在着只在ISDN网络中操作的路由器，也存在着只为模仿租用线路的模型，等等。

路由器性能的需求越少，它根据LAN的第一个路由器（以及桥）的经典设计而设计的可能性

越高——即，基于单个中央处理单元以及不具备专属处理器端口的基础。这种设计相对便宜得多。但是，它的性能完全依赖于处理器的性能并且不能通过端口数的增长扩大规模。

软件路由器（software router）是某些通用操作系统的特殊软件模块，无论是UNIX或Windows系统。

只有这种高速技术，如ATM、SONET/SDN、DWDM的到来增加了对路由器需求的性能。这导致高端类路由器的代表进化到具有交换结构的多处理器设计，在LAN交换机上已经得到了测试成功。

LAN路由器（LAN router）（第3层交换机）旨在将大型LAN划分为子网。这是路由器的特殊分类，通常没有WAN接口。

此类的很多路由器起源自LAN交换机，这赋予了它们该名称——第三层交换机。第三层交换机执行所有的路由器功能，除此之外它们还能充当普通的LAN交换机（即第2层交换机），操作模式（交换机或路由器）视配置参数而定。

这种设备也可以操作于联合模式，其中第三层交换机的多个端口拥有同样的IP地址（图20-12），此时，在属于同一网络的多个端口之中的分组传输在数据链路层以交换模式执行（例如，基于MAC地址）。如果端口属于不同的IP网络，那么交换机执行网间路由。分组传输模式的选择由端口的IP地址来决定，因此，也就是由计算机的配置来决定。

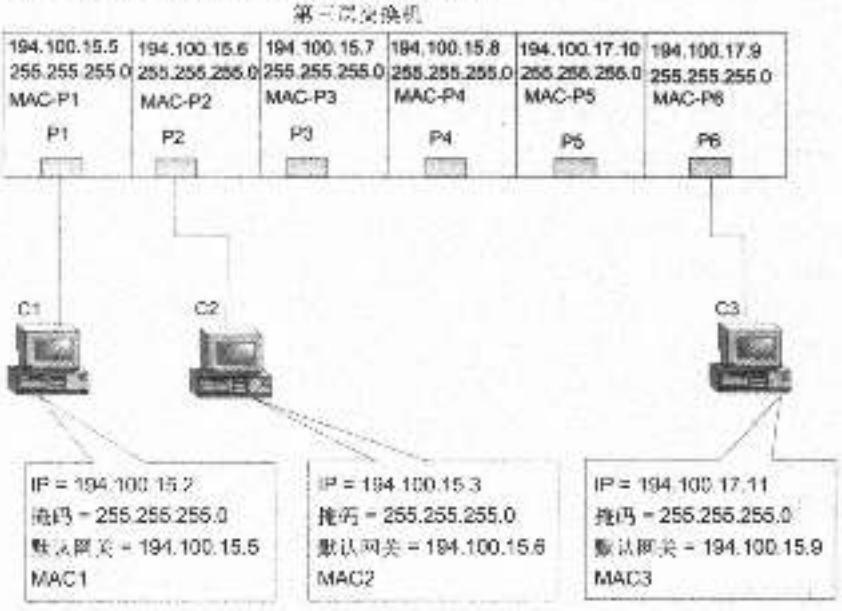


图20-12　第三层交换机的联合操作模式

示例　例如，如果两个计算机（图20-12中的C1和C2）拥有属于同一个网络的地址，那么交换信息时它们不会传递分组给默认路由器，取而代之的是，它们会利用ARP来发现目的计算机的MAC地址，假设计算机C1需要传递分组给计算机C2。第三层交换机将来自计算机C1的ARP请求帧以及广播MAC地址传递给属于同一IP网络的所有端口（如端口P1，P2，P3，P4）。计算机C2识别出该请求中的IP地址（194.100.15.3），并做出响应，即发送帧给计算机C1的目的MAC地址（MAC1），携带它自己MAC地址的信息（MAC2）。之后，计算机C1向计算机C2发送IP分组，将MAC2目的地址封装在帧里。基于第二层转发表，第三层交换机根据网桥算法将此帧从端口P1传递到端口P2。第三层交换机也将以类似方式操作。

当计算机属于不同IP网络时，发送端计算机的行为显示了分组传递到第三层交换机的

方法。比如，如果计算机C1发送分组给计算机C3，后者位于不同的子网中，它必须传递分组给默认路由器而不是试图利用ARP找到目的计算机的MAC地址。因此，计算机C1发送ARP请求来获取已知默认路由器的MAC地址。这种情况下，这是端口P1，IP地址为IP-R1。收到端口P1的MAC地址之后（MAC-P1），计算机C1发送预前往计算机C3（即目的地址为194.100.17.11）的IP分组给那个端口。此IP分组被封装在目的地址为MAC-P1的以太网帧中。在收到携带自己MAC地址的帧之后，第三层交换机以路由方式而不是交换方式来处理它。

第三层交换机，作为将独立VLAN连入IP内部网的主要设备类型，支持VLAN技术。通常，每个VLAN都被分配一个IP网络号，这样VLAN内的分组传输建立在MAC地址基础上，VLAN之间的数据传输建立在IP地址基础上。在图20-12中显示的例子中，端口P1～P4属于VLAN1，端口P5和P6则属于VLAN2。

小结

- IP路由器允许基于属性过滤用户数据流，这些属性包括源地址、目的地址、IP分组携带的协议类型以及UDP/TCP端口号。路由器的这个特征广泛用于保护网络免受攻击并且限制合法用户的访问。
- 路由广告的过滤从总体上保证了对网络连接的控制，并阻止特定链路的记录出现在路由表中。
- IP路由器一直以来支持不同的QoS机制，包括优先级和加权队列、流量监控以及TCP流量的反馈。但是，只有在20世纪90年代中期对IP QoS标准的研发才启动，此时需要在因特网上传输延迟-敏感的流量。
- 如今，存在着两个系统的IP QoS标准，集成服务和区分服务。第一个系统，利用RSVP信令协议来预留路由器资源，确保微流传输的质量。集成服务解决方法的缺陷是主干路由器上的大量负载，它必须存储数以千计的用户流的状态信息。
- 区分服务技术使用的是一种聚合解决方案，它确保少量流量分类的QoS。这大大减少了路由器上的负载。而且，区分服务基于每跳行为模型，每个路由器决定它必须为每个分类贡献哪些资源。这同时简化了路由器的操作，提供了每个路由器决定为每个流量分类分配哪些资源的可能性。但是，简化的区分服务解决方法降低了QoS的承诺水平——即，增加了QoS超出客户端所需限制发生的可能性。
- 典型的路由器是一个可编程的计算设备，它在特殊OS控制下操作，此OS经过优化以执行构建路由表并在这些表的基础上转发分组。
- 路由器通常构建在多处理器设计上。更常见的情况下，使用对称多处理、非对称多处理或它们的组合。与分组处理相关的多数常规操作由专用处理器按程序执行或者在硬件层实现(LIC/ASIC)。更高层的操作则由通用处理器按程序执行。
- 路由器可以利用不同方法分类：它们可以被分为主干和边缘路由器（根据它们与网络边界的相对位置）、载体和企业路由器（根据拥有该网络的公司类型）。企业范围网络中操作的路由器通常被划分为企业路由器（操作在公司的核心网络）、区域部门路由器以及远程办公室路由器。同时存在着LAN的特定分类路由器。它们不支持WAN接口，通常被称为第三层交换机。
- 网络地址转换（NAT）技术使得公司可以解决IP地址短缺问题并能通过隐藏主机地址来提升网络安全性。这通过使用内部网中的私有地址来实现。当分组离开内部网的限制范围时，私有地址即被翻译成全局IP地址。
- 传统NAT更进一步划分为只使用IP地址映射的基本NAT技术以及所谓的传输标识符也被用来转换的网络地址端口转换技术（NAPT）。更常见的情况下，TCP/UDP端口号被用于此目的。

复习题

1. 路由器流量过滤时可以使用哪些分组参数：
   A. 源IP地址
   B. IP分组携带的协议
   C. UDP/TCP端口号
   D. 前一个分组的源IP地址
2. 路由广告与用户流量过滤的结果之间有什么区别？
3. 集成服务技术的名字中“集成”是什么意思？
4. 什么参数可以限制使用令牌桶算法定义的输入分组流发生突发？
5. 为什么RED方法中分组丢弃率依赖于平均队列长度而不是当前队列长度？
6. 为什么RED机制不适用于UDP流量？
7. 请解释利用RSVP进行路由资源预留的主要步骤。
8. 集成服务技术的主要局限是什么：
   A. 它不能应用于多播寻址
   B. 路由器必须存储每个链路的状态信息
   C. 终端节点必须阶段性地更新预留
9. 为什么区分服务技术不使用信令协议？
10. EF和AF服务间的区别在哪里？
11. 区分服务技术的什么特殊特性使得它在通信载体中很受欢迎：
    A. 它可以在通信载体网络限制范围内实施，独立于其他通信载体网络
    B. 路由器对流量分类进行操作，不给路由器造成大量负载
    C. 它确保了QoS计算的自动化
12. NAT的主要目标是什么？
    A. 阻碍DoS攻击
    B. 解决IPv4中IP地址短缺的问题
    C. 保护公司的内部地址空间
13. 哪个附加分组属性被用于NAT中以实现内部地址集到单个全局地址的映射？
14. 填充NAT表中的“分配端口”栏。

私有地址	发送端端口	全局地址	分配端口
10.0.25.1	1035	193.55.13.79	
10.0.25.2	1035	193.55.13.79	
10.0.25.3	1035	193.55.13.79	
10.0.25.2	1047	193.55.13.79	
10.0.25.1	1047	193.55.13.79	

15. 列出路由器架构的主要变化形式。
16. 为路由器分类的标准是什么？
17. 第三层交换机的特定特性是什么？

练习题

1. 为连接公司和因特网的Cisco路由器构建一个访问列表。该访问列表必须确保下列描述的内容：
   A. 网络194.100.12.0/24的用户，除了用户194.100.12.25以外，只与网络132.22.0.0/16和

201.17.200.0/24中的节点通信。这些用户不允许与因特网交换信息。

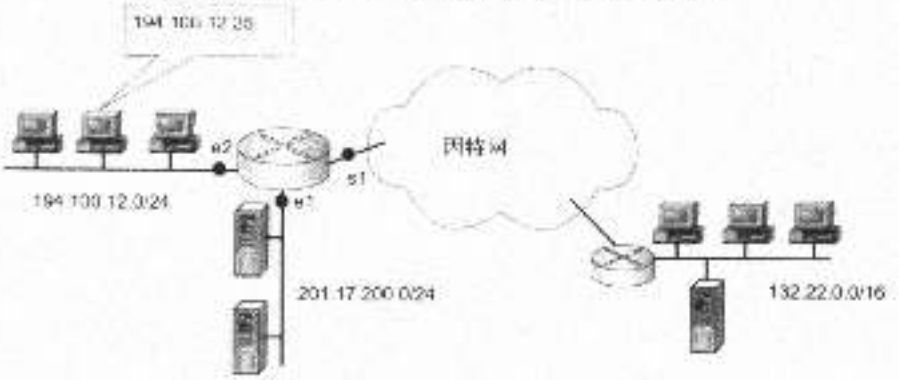


图20-13 使用路由器进行流量过滤

B. 用户194.100.12.25必须具备无限数据交换的能力。

C. 因特网只利用ftp和http访问网络201.17.200.0/24的服务器，而不允许利用ICMP来访问。

2. 在描述NAT技术时，我们简化了这个模式。特别地，我们并没有考虑在ICMP错误报文抵达内部网时可能引起的问题。请对该算法作出修改，使它可以在ICMP报文抵达外部接口时可以被NAT协议使用。

提示 在传递ICMP报文之前，ICMP必须更正IP头部以及ICMP数据字段。

第五部分 广域网

本书前半部分介绍的IP技术允许创建多种类型的互联网络，包括局域网（LAN）和广域网（WAN）。除了IP以外，还有其他技术专门用于创建广域网络。目前广泛使用的有：只从历史角度看起来很有趣的X.25网络，以及帧中继和ATM网络。这些技术有一个共同的特点——它们都是基于虚电路的。我们在第一部分提到，这种技术是数据报的分组转发技术的另外一种实现方式，以太网络和IP网络正是基于此。这两种数据传输原理的竞争自从第一个包交换网络的出现就已经存在了。

自互联网革命以来，商业广域网（WAN）使一直倾向于虚电路技术。选择这个方式的原因是相比数据报的分组转发技术，对于网络用户间的连接以及信息在网络节点间的路由而言，该技术确保了更高的控制能力。结果是，传播媒体能够理性的控制用户申请的服务之间的资源分配。利用虚电路也能够提供更高的QoS。很显然，这种方式也有代价，那就是为了建立每个虚拟连接，要付出更多的开销（时间上和金钱上）。网络节点之间的数据报通讯技术正好相反，在任意两个网络节点间建立连接非常简单。但是它限制了操作者控制用户间资源分配的能力。现代网络可以在这两类方式中寻找一个折衷方案。在由广域网组成的互联网环境中，大部分的代理网络都是基于虚电路操作的，就是说都是帧中继和ATM网络。这些网络通过IP数据报协议互联。这样的多层次方式能够构建出理想的广域网。但是，广域网组织会变得很复杂。另外某些功能也是部分重复的。例如，在ATM网络的IP层运作路由协议。

为了使数据报技术和虚电路技术更好地结合，多协议标记交换技术（MPLS）出现了。MPLS使用TCP/IP的路由协议来探测网络拓扑和寻找合理的路由，另一方面，它使用虚电路技术来转发网络分组。

确保网络主干的高速访问是当务之急。对于用户到最近的通信载体中心的亿万次连接，很有必要提高访问频率。因此，传统中主要基于光纤并需要安装新线路的主干方案，对于多数家庭用户来说成本过高。利用已有线路结构的技术，例如基于本地环和已有电话线路的非对称数字用户环线（ADSL），又或者使用电视网络的电缆调制解调器，是更有效的。固定或者移动无线网络访问也是一种解决方案。

所有这些技术和方法都综合为第五部分。主要针对不同类型的网络服务：信息服务、安全服务和网络管理服务。

第21章 虚电路WAN

21.1 引言

几种广域网（WAN）技术比如X.25、帧中继和ATM主要的区别在于它们的功能特性。同时，它们都使用了虚电路技术，即面向连接通信。在第3章，我们讨论了这种机制的一些基础特性。在这里，我们更多地考虑以上几种技术在部署和实现特别功能的细节方面。

虚电路技术是逐渐演化而来的。X.25技术出现于计算机网络时代的初期，实际上与APPANET项目同步出现，后者正是互联网和IP数据报协议的起源。X.25网络使用虚电路技术来确保可靠的数据传输，这也是在20世纪70年代和20世纪80年代这种技术变得流行的主要原因。这是因为当时的许多连接都是模拟线路，无法自动确保可靠的数字信号传输。因此，X.25网络能够修复损坏的和丢失数据分组的能力在当时显得非常可贵。

伴随着20世纪80年代中期快速稳定的数字连接的出现，X.25确保可靠数据传输的特性开始显得多余。该技术革命导致了新广域网技术的出现——帧中继。该技术在广域网中和以太网在局域网中扮演的角色相同：仅仅携带能够把帧传递到目的地的最少功能集合。去除掉很多当代电信不需要的功能不是帧中继和X.25技术的唯一区别。新技术也添加了一个非常重要的特性，即对于有弹性流量的质量控制服务。刚开始的时候，帧中继标准的开发者并没有打算支持延迟敏感的流量。因此，延迟度和抖动并没有被包括在对用户开放的参数中。然而帧中继网络中的高质量话音传输是可能的，因为有支持传输优先级的交换机存在。

ATM技术为用户提供了通用的综合传输服务。相对于X.25和帧中继，ATM最初的设计目的是面向所有现存传输类型的：计算机数据、语音、视频、对象控制等等。在这种技术中被称做单元、固定大小的小型帧可以最小化实时传输的延迟。ATM还能把分离的虚电路聚合为虚拟路径，从而提高了它的可延拓性。然而，高质量的服务同时意味着高代价，因为ATM网络的技术更加复杂和昂贵。另外一个阻止ATM变成一个通用流量载体的原因是在超高频的传输率下，比如2.5和10Gb/s，处理传输单元变得很困难。无论怎样，ATM还是很流行的，目前还没有其他技术拥有类似的质量控制服务和流量工程特性。

21.2 虚电路技术

有两种类型的虚电路：

- **交换虚电路**（switched virtual circuit，SVC）SVC是由网络终端节点使用一种自动过程主动创建的。
- **永久虚电路**（permanent virtual circuit，PVC）PVC是由人工配置的网络路由器预先创建的。这个过程是由网络管理员完成的，可能使用集中化的网络管理系统和一些控制协议。（通常情况下，这是一种专用协议）

类似SVC/PVC这样的简写也经常被称做交换/永久虚拟通道或者交换/永久虚拟连接。

让我们从SVC的创建过程开始。

21.2.1 交换虚电路

SVC创建原理和在第3章我们提到过的电话网络中建立一个连接的过程相类似。在电话网络中，

用来创建该过程的协议被称为信令协议。因此，在分组交换网络中用来建立虚电路的协议也常被称为同样的名字。

虚电路的创建也需要网络交换机中出现与数据报网络中的路由表类似的路由表，如IP网络。用来创建这类表的方法——无论是人工还是自动——并不重要。如图21-1中的表所示。

图21-1表示了创建一个通过网络来连接节点（N1，A1）和（N2，A2）的虚电路的过程，该例子里分别由两个路由器代表——S1和S2。路由表指定了目的网络的地址。该过程发生在三个阶段，在图示中和下面的描述中使用数字标明。

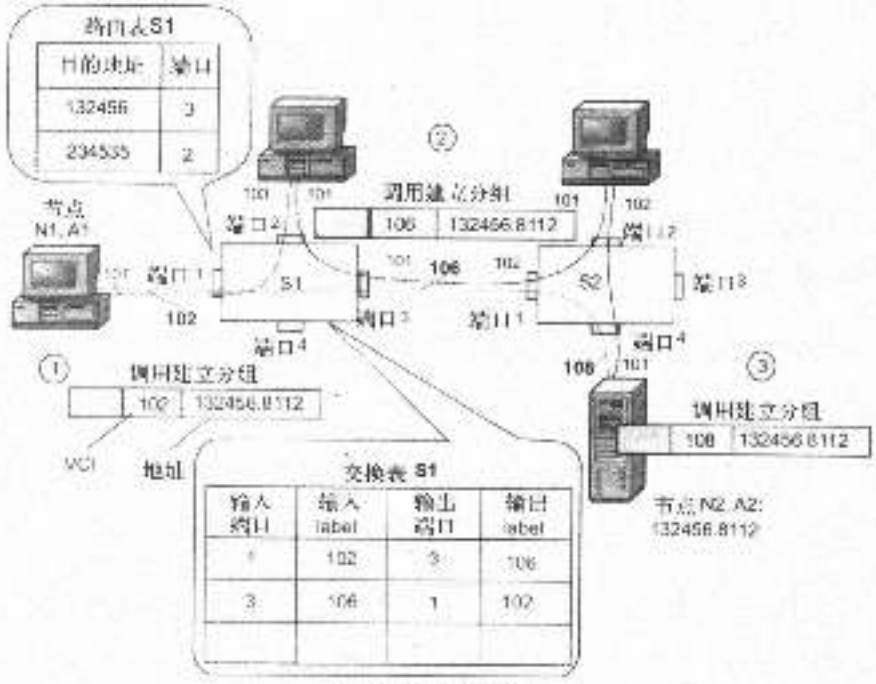


图21-1 建立虚电路

1）虚电路的创建始于初始节点（N1，A1）产生了一个特定分组——一个建立逻辑连接到节点N2，A2的请求。在这个一般性例子中，该请求被称做**呼叫建立**（call setup）（在一些特定的信令协议，例如帧中继的Q.933协议中和ATM的Q.2931协议中有同样的名称）。该请求包含一系列的值——目的节点地址和虚拟通道标识符（VCI）的初始值。在这个例子中，最初的呼叫建立请求使用了如下的格式：

（102，132456.8112）

这里102是VCI的初始值，132456.8112是目的地址。其中高位部分是子网号，低位部分则是节点号[⊖]。

将数字102赋予给虚电路，还有一个本地的值用来表示该连接是从于计算机哪一个端口建立的。因为已经有一个虚电路通过了该端口（电路数字101），在终端节点运行的信令协议软件从允许范围内选择了第一个可用的数字（也就是，还没有被使用的数字）。这种方法保证了每个端口上虚电路标识的唯一性。

发起者必须选择合适的网络交换机来建立一个虚电路。这个选择可以根据发送端节点的路由

⊖ 此示例使用3个字节的子网地址以及2个字节的终端节点地址。实际上，基于虚电路的WAN通常使用更长的地址。

表来决定；但是如果该节点是像图21-1那样通过一个单一端口连到网络的，那么在终端节点上就没有必要存在一个路由表。当呼叫建立分组到达交换机S1的端口1的缓冲后，根据它的目标地址和在**路由表**（routing table）中找到的值来处理它。这个携带了目的网络地址132456的记录指定了该网络分组应该被送到端口3。

说明　注意和IP网络的路由表相反，图21-1中的路由表并不包含下一个路由器的地址。这是因为广域网路由器并不支持多个连接，而是总被物理“点到点”连接起来的。因此，输出端口号明确地指明了下一个路由器。

2）决定了呼叫建立分组的输出端口号后，交换机S1生成了下一个VCI号码——106。选择该数字的原因是它是第一个可用的数字，并且在该本地网络的限制里，它唯一地表示了正在建立的虚电路。我们考虑到这个情形是因为早就提到，VCI编号本质上就是本地的。改变VCI值后，呼叫建立网络分组采用了格式（106,132456.8112），然后被交换机S1的端口3输出到交换机S2的端口1。

在分组转发的同时，交换机创建了**交换表**（switching table）（不要和上文提到的*路由表*（*routing table*）混淆）。当建立虚电路时，要通过该表传输用户数据，这时就不用再指明目的地址了。

每个交换表的记录包含了以下四个主要字段：

- 输入端口号
- 到达输入端口的分组中的输入标签（VCI）
- 输出端口号
- 通过输出端口传输的分组中的输出标签（VCI）

交换表中“1-102-3-106”记录表示所有到达端口1并且携带VCI 102的网络分组都会被转发到端口3，并且VCI字段会被改写为新值——106。

虚电路既可以是双向的，也可以是单向的。在前面的例子里，电路是双向的；因此，交换机在交换表里面创建了另外一个记录——分组反向转发，从节点N2，A2到节点N1，A1。该记录镜像了第一个，因此具有VCI标签106并且到达交换机S1端口3的分组将具有VCI初始值102。结果就是，节点N1，A1能够正确识别到达分组所属于的虚电路，尽管其编号在网络分组传输过程中发生了很多变化。

3）交换机S2继续建立虚电路的过程。为了达到这个目的，它在请求中指定的目的地址。使用交换机的路由表（图21-1没有包含），交换机能够判定传递分组须经过的输出端口。在该操作同时，VCI字段被更新了。该例中，交换机给呼叫建立分组指定了VCI值108。因此，呼叫建立分组到达目的节点的格式是：（108，132456.8112）。收到该请求后，终端节点可以接受或者拒绝它。如果是肯定的决定，终端节点通过向发起者发送连接服务分组。该分组会通过交换表中的“镜像”记录来传输。连接分组向所有的交换机和发起者确认一个虚电路建立。

发起者收到连接确认之后，终端节点可以开始利用虚电路来发送用户数据。N1,A1终端发送的信元在VCI基础上转发，通常是很短的。例如，在X.25技术中，它的长度只有1.5个字节。相比之下，在X.26网络中目的地址可以有16个字节。

从原理上，SVC技术使用两种网络操作模式：

- 建立SVC的时候，建立连接的请求是通过标准路由模式的网络传输的。该模式使用整个网络全局化的目的地址，要求网络拓扑的完整信息。这意味着建立虚电路的协议（信令协议）运作在OSI模型的网络层。
- 连接建立之后，网络是基于本地地址和本地转发表运作的，这样这种模式能够被划分到OSI模型的数据链路层。通信设备被称为交换机（这是该层次设备的一个标准名）。

21.2.2 永久虚电路

PVC模式不允许终端节点动态地创建虚电路。取而代之，网络管理员预先手工创建交换表。管理员可以本地完成这项工作——例如，通过笔记本电脑作为一个虚拟终端，使用RS-232接口连接到路由器。当然，这种配置路由表的方式不是最方便的，尤其是对于像WAN这样的分布式系统。因此，通常管理员依赖于网络管理系统。这类网络管理系统通过某一种管理协议与网络路由器进行通信，这在第22章："简单网络管理协议或者公共管理信息协议"中提到。管理员提供给网络管理系统配置信息，指定建立虚电路必须通过的节点。网络管理系统自动选择VCI需要的值和创建交换表中的记录来与网络路由器通信。不幸的是，标准网络管理协议的使用无法确保网络管理系统之间的兼容性。这是因为网络管理系统是由不同制造商分别部署的复杂的应用程序。

因此，只有基于同一个制造商提供的设备，网络段内自动实现PVC才是可能的。至于在网络边界上的PVC分段，它们必须被手工"粘合"。

路由表在PVC中是不必要的，因为路径往往是由管理员进行选择的。

为了使新创建的虚电路发挥作用，管理员需要在终端节点上输入对应的编号。注意该编号对于两端来说都是不同的。例如，如果图21-1中的虚电路是被永久创建的，那么计算机N1，A1的管理员必须输入标签102，而对于计算机N2，A2——标签为108。

假如虚电路技术只支持PVC模式，那么用户可以认为它是一种专门的数据链路技术。帧中继就是很好的例子，因为很长时间只有PVC模式存在。因此，它被归类为数据链路技术是非常合理的。尽管今天帧中继网络两种模式都支持，考虑到数据转发模式，它仍然被归类为数据链路技术。与帧中继恰恰相反，ATM技术一开始就支持两种模式。然而它通常还是因为同样的原因被分类为第2层技术。

21.2.3 与数据报技术的比较

与数据报技术相比，虚电路技术有优有劣。

与数据报协议不同，例如IP，虚电路协议要求在用户数据交换之前就建立连接。这导致了数据传输开始前额外的延迟。这种延迟当传输在传输少量数据——所谓的**短期数据流**（*short-term data flow*）时变得特别明显，因为建立虚电路的时间相对于数据传输的时间显得比例非常高。

数据报网络由于没有连接建立阶段，因此传输短期数据流（short-term data flow）效率更高。支持虚电路的网络更适合用于传输长期数据流。

尽管如此，不要忘记建立虚电路的时间也被整个数据流的快速传送所弥补。在支持虚电路技术的网络里路由快速的原因主要因为以下两点：

- 首先，转发特定网络分组的决定能够更快完成，因为交换表比较小；
- 其次，在网络分组里的控制信息量更少。在广域网中，终端节点的地址一般都比较长；通常，它们都是14～15个十进制数字。因此，它们需要在分组的头部中占用20个字节。与这种情形不同，虚电路的数字一般不超过10～12字节。

PVC模式从性能角度上来说，是最优的。分组路由大部分工作已经被网络管理员完成。在这种情况下，路由器只根据预先创建的交换表来尽可能快地转发网络分组。一个PVC线路在很多方面都很像一个租借线路，因为没有必要来创建或者终止连接，并且分组交换可以按需开始。PVC和租借线路的区别是用户没有带宽保证。使用PVC的主要好处是大大便宜于使用租借线路，因为用户和其他网络用户共享带宽。

PVC对于传输包含大量网络用户的独立数据流组成的**聚合通信流**（aggregate traffic flow）比较有效率。在这种情况下，虚电路创建于有聚合通信流所在的骨干部分。它连接两个边缘设备而

不是终端用户，比如，两个通信载体的边界路由器。根据大数据定理，聚合通信流具备高度稳定性的特征，因此不用动态创建SVC。在这种情况下，使用永久电路将更有效，况且它可以使用流量工程来很好地规划。

数据报网络另外的一个优点是能够快速适应拓扑变化，比如由路由器或通信链路失效导致的。当这类情况发生时，分组可以很简单地被发送到另外一条路径。当然，在这里也需要考虑到在路由表内建立新配置的延迟。另一方面，当通信链路失效或者虚电路上的路由器出错时，这个虚电路连接就中止了，需要创建一个新的虚电路来绕过发生故障部分。

21.3　X.25网络

21.3.1　X.25网络的结构和目的

X.25规范“为在网络分组模式下使用专有电路连接到公用数据网络的终端而设计，数据电路——终端设备（DTE）和数据电路—终端设备（DCE）之间的接口”，是1974年由CCITT（现在的ITU-T）开发的。自那以后，其被修改过数次。

从名称可以看出，该标准并不描述X.25网络的内部架构。相反，它仅仅定义了以广域网为代表的网络用户界面，该界面被称为**用户—网络界面**（user-to-network interface）（UNI）。网络的内部架构是任意的，可以由通信载体来选定。实际上，广域网路由器使用类似UNI的协议来通信。为了组织由不同通信载体所有的网络之间的通信，有必要开发一个**网络—网络接口**（network-to-network interface）（NNI），通常它是UNI的一个变更版本。

X.25网络有一些很重要的特性，使它和别的技术区别开来：

- X.25最适合传输20世纪70年代到80年代间典型低密度流量，不太适合像LAN流量这样更高的要求。
- 特殊设备的存在，**网络分组编译器—解编译器**（the packet assembler-disassembler，PAD），是为了编译几种从字符终端到一般计算机的低速的启动—停止字符流。PAD的出现是X.25技术划时代的标志，因为在20世纪70年代早期，字符终端除了执行一些很原始的操作，什么也干不了，其他所有的操作，和数据链路还有网络层功能相关的操作都是由PAD执行的。

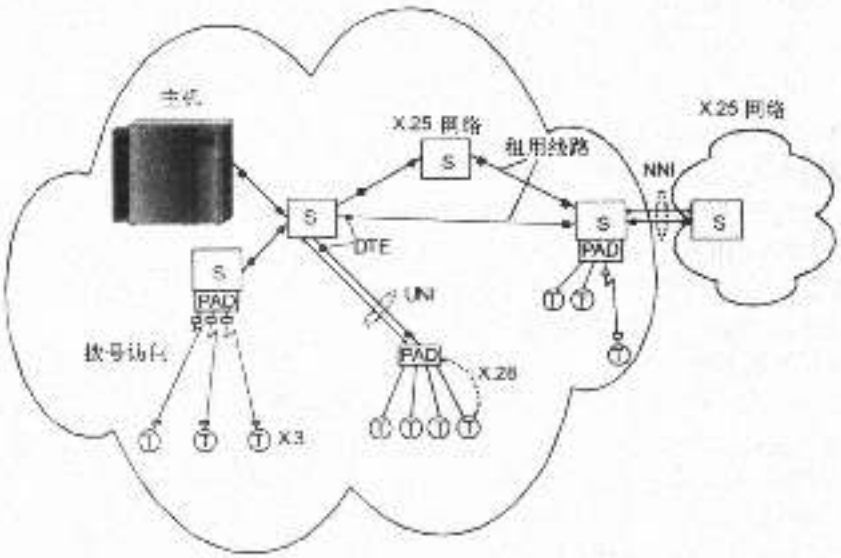


图21-2　X.25网络的结构

- X.25的三层协议栈的存在，它在数据链路层和网络层使用面向连接协议来控制数据流以及更正错误。这样确保可靠数据传输的冗余功能，是针对具有BER特征的10^{-3}到10^{-4}个不可靠通信线路来设计的。
- 面向所有网络节点传输协议的统一栈。因为网络层本意是和数据链路层的协议一起工作的，和IP对比，它不能互联异构网络。

X.25网络由高速租用线路连接、地理位置上分散的路由器组成。租用线路可以是数字的或者模拟的。

PAD可以是远程的或者内建的。一般来说，一个内建的PAD被安置在交换机架上。终端通过使用调制解调器和电话线访问内建的PAD。远程PAD是一个通过专用X.25连接到交换机的独立的小设备。终端通过一个异步接口连接到远程PAD，比如RS-232C接口。单个PAD通常为8、16或者24个异步接口提供访问。终端不会被赋予X.25网络的终端节点地址。地址被授给PAD的端口，这些端口通过租用链路接入X.25网络分组交换机。

计算机和局域网通常使用X.25适配器或者在界面上支持X.25协议的路由器直接连接到X.25网络。

21.3.2 X.25网络寻址

如果X.25网络和外部世界没有连接，它能够使用任意长度的地址（在地址栏格式允许的范围内），可以被赋予任何值。在X.25网络分组中最大的允许长度是16字节。

X.121 CCITT 标准定义了在公共数据网络中的地址取值的国际化系统。如果X.25网络需要和别的X.25系统交换数据，它必须遵循X.121寻址标准。

X.121地址，又称为**国际数据号**（international data number，IDNs），有着不同的长度，可达14位十进制数字。

- IDN的头4位数字是**数据网络标识代码**（data network identification code，DNICs）。DNICs被分为两部分：
  - 第一部分（三个十进制数字）定义了该网络所在的国家
  - 第二部分（一个十进制数字）是该X.25网络在本国的数字代号

因此，每个国家只拥有10个X.25网络。如果需要更多，需要给该国家赋予多个代码。

- 所有其他的数字表示**国家终端号**（national terminal number），用来在X.25网络中标识一个特定的DTE设备。

21.3.3 X.25网络协议栈

X.25网络标准定义了三层协议（图21-3）。

- 在*物理层*（*physical layer*），对于数据传输设备有两个**同步接口**（synchronous interfaces），X.21和X.21bis。当租用线路为数字的时，就是连接到DSU/ CSU的；如果是模拟线路，就是到同步调制解调器。通信链路的物理层协议没有预先定义，这样就可以针对不同标准使用不同的连接。

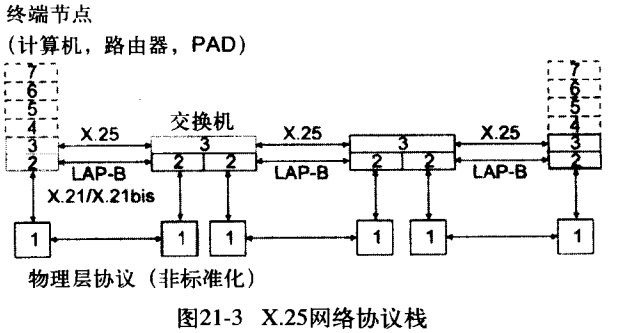


图21-3 X.25网络协议栈

- 在*数据链路层*（*data link layer*），**平衡的链路访问流程**（link access procedure-balanced）（LAP-B）用来确保数据线路上出错

后的自动重传。该协议确保一个平衡的操作模式，即两个参加连接的节点是对等的。根据LAP-B，连接是在DTE（计算机，IP路由器，或者IPX路由器）和网络交换机之间建立的。尽管协议本身没有明确指出，LAP-B也常被用来在数据链路层中直接连接的设备之间建立连接。LAP-B是一种面向连接的协议，使用滑动窗口协议来确保两台直接连接设备间的可靠帧传输。和TCP-IP相比，LAP-B使用该算法更加简化的实现方式。此时传输帧被标上了编号，而不是设为字节。窗口不能动态变化，而是具有固定大小为8或者128的帧。LAP-B属于高层数据链路控制（HDLC）协议家族，这在第22章中会更加详细地介绍。

- 在网络层（*network layer*）（该标准使用网络分组层（*packet layer*）术语代替网络层（network layer）），该标准定义了在终端设备和数据传输网络间分组交换的X.25/3协议。LAP-B连接确保一对邻接节点间的可靠传输，但是不能在终端节点间交换信息。为了建立一个虚拟的端对端连接，需要使用X.25/3协议。

关于X.25/3协议的操作以下将列出更多的细节。其主要的功能有：

- 网络分组路由
- 在网络用户间建立和终止虚电路
- 控制网络分组流

根据该协议，终点节点发送封装为LAP-B的呼叫请求分组。

注释　相对于其他基于虚电路技术的网络，X.25网络不使用独立的信令协议。但必要时，X.25/3协议靠切换特殊操作模式来实现这个功能。

呼叫请求分组（*call request packet*）使用X.121格式来指明源和目标地址。网络交换机收到呼叫请求分组后，根据路由表做出路由，从而创建一个虚电路。路由协议并不是为X.25网络定义的；因此，这里的路由表总是由人工创建。

因为呼叫请求分组根据路由沿着交换机逐个传输，这使得它们在交换表中生成新的记录并为之分配新的标签值。于是，一条新的虚电路产生了。虚电路号的初始值是用户在网络分组中的逻辑渠道编号（*logical channel number*）（*LCN*）指明的，该字段为我们描述建立虚电路原理时VCI字段的模拟。

建立了虚电路之后，终端节点使用另一种格式来交换网络分组——数据分组。在数据分组中，源和目标地址都没有指明，只有*LCN*标签保留了所有的地址信息。

X.25技术与帧中继技术和ATM的区别是，它是一个网络层的技术。实际上，一个虚电路建立后，数据传输由网络层协议执行，而不是数据链路层协议。

21.4　帧中继网

相对于X.25网络，帧中继是一种更新的技术，能够更好地适应计算机网络中典型的突发数据流。这种优势只有在通信链路的质量能与局域网中的相当时，才体现出来。对于广域网来说，这样的质量只有使用光纤才有可能实现。

首先，CCITT将帧中继技术作为一种综合服务数字网络（ISDN）服务（RFC2955）来标准化。ISDN技术首先是为了实现一个全球化的网络来设计的，该网络旨在提供各种电话和数据传输服务。很不幸，这个雄伟的项目没能达到最初的目标。下一代技术在其他技术基础上创建起来，如IP。尽管如此，几个同样重要的目标在执行该计划过程中被实现了。这些成就的列表中包含了对帧中继技术的开发，该技术后来变成了独立于ISDN[⊖]的技术。

⊖ ISDN技术将在第22章中简单介绍。

在1988年发布的I.122标准中，该服务被列为ISDN网络分组模式的附加服务。尽管如此，在1992或1993年，当修改这些标准时，两项新的服务被定义了——帧中继（frame relay）和帧交换（frame switching）。这两种服务的区别是，帧交换确保了帧的传递，而帧中继只提供尽力服务。

虽然简单，但在光纤通讯链路中非常有效，帧中继技术立刻吸引了大部分参与标准化的电信服务商和组织。

除了CCITT（ITU-T）之外，还有其他的组织和投参与了该技术的标准化工作。这些组织包括了帧中继论坛（FRF）和T1S1 ANSI委员会。而帧交换技术，一直保留在标准的状态，没有广泛投入应用。

ITU-A和FRF开发的帧中继标准，定义了两类虚电路——永久的（PVC）和交换的（SVC）。根据用户的不同需求，永久电路适应可靠传输的连接，而那种一个月实际使用只有几个小时的连接来说，SVC更加合适。

尽管如此，帧中继设备的制造商和帧中继网服务的提供商已开始只支持PVC，很显然，这是对该技术的一种简化。SVC的设备在市场上出现时间相对晚一些；因此，帧中继很多时候只和PVC有关。

21.4.1 帧中继协议栈

帧中继协议栈比X.25技术栈简单。帧中继开发者考虑到了1980年出现的高质量的光纤通信链路，他们认为不在协议栈内包括可靠性功能是可行的。使用这样的通信链路，很少发生错误。如果非开这些错误发生的极低概率，仍然发生了问题，帧中继会忽视它，把这些丢失或者受损网络分组的恢复交给具有相关功能的高层次协议，例如TCP-IP。

由于协议的低开销性，帧中继有能力确保大的带宽以及低的延迟速率。

注释 和帧中继同时发展的帧交换技术与X.25技术类似，主要为了确保数据链路层上可靠的传输。该技术可以在没有满意的通信链路质量时，或者数据链路层因为某些原因要求确保可靠传输时使用。然而实际上，帧交换技术并没有得到广泛应用。尽管如此，在这里还是简要描述一下，因为帧中继协议栈就是考虑到该技术的存在才被创建的。

图21-4按照ITU-T标准的描述，显示了帧中继和帧交换的协议栈。控制平面协议执行所有建立虚拟连接的操作，而数据平面协议则使用已经建立的虚电路传输帧。

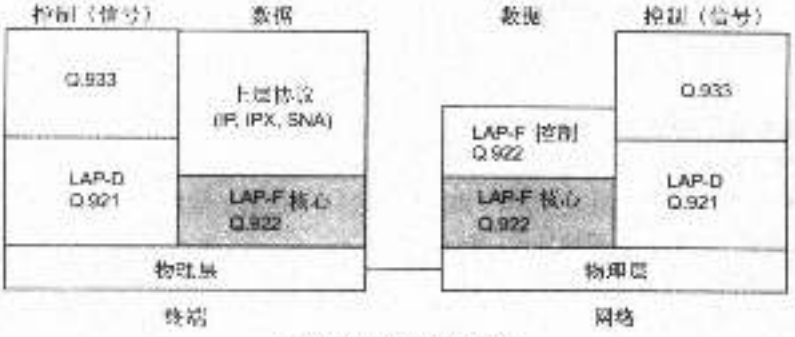


图21-4 帧中继协议栈

在ITU-T标准中被称为Q.922的*帧模式承载服务的链路访问过程*（*link access procedure for frame mode bearer service*，LAP-F）。在帧中继网中运行于数据链路层，该协议有两个版本。

- LAP-F 核心是一个在所有帧中继网中运行的核心部分，在图21-4中灰色的是LAP-F核心，提供了供建立帧中继网的最小工具集。这样的网络仅能提供PVC服务。

• LAP-F控制协议在需要提供帧交换服务的网络中运行。

LAP-F核心和LAP-F控制协议都是数据链路层协议，它们确保了相邻交换机之间的帧传输。

在物理层（*physical layer*），帧中继网能适应PDH/SDH链路或者ISDN链路。

现在我们来考虑负责建立动态SVC的控制层。要建立SVC模式，网络交换机需要支持两种控制协议——数据链路层的**链路访问过程D**（link access procedure D）（LAP-D或者Q.921）和网络层的Q.933。帧中继中的LAP-D确保相邻交换机间的可靠帧传输。

Q.933协议使用其间建立了虚电路的终端节点地址，通常这些地址根据E.164规范产生的电话号码指定。一个地址包含15位十进制数字，与一般电话号码类似，分为国家代码（*country code*）字段（1-3数字0），城市代码（*city code*）字段，用户号（*subscriber number*）字段。ISDN地址包含长度可达40位数字的**子地址**（subaddress）。当用户有多个同类设备时，子地址用来标识紧跟DTE的终端设备。

在帧中继技术中没有定义自动创建路由表的协议，因此，可以使用设备制造商的专利协议。另一个方法是，手工创建路由表。

注释　帧中继技术相对于X.25的最大优势是，在建立虚拟连接之后，仅仅使用数据链路协议传输帧。在X.25网络中，用户数据在建立连接后，由数据链路层和网络层协议共同传输的。使用这种方式可以减少帧中继技术传输局域网分组的负担，因为所有数据分组都被封装到数据链路层帧，而不是封装到X.25网络的网络层的分组中。

帧中继技术通常被划分为数据链路层的技术，其大部分努力花费在传输用户数据的过程上。建立虚电路的过程使用网络层协议来完成。

帧中继虚电路能用于传输不同协议的数据。RFC1490指明了方法，可以将网络协议分组诸如IP、IPX和SNA数据封装为帧中继的帧。

LAP-F的结构如图21-5所示。

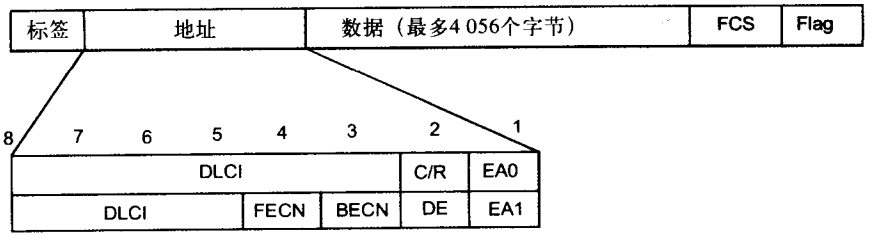


图21-5　LAP-F帧格式

数据链路连接标识符（data link connection identifier，DLCI）字段为10位，这样能够使用最多1024个虚连接。DLCI字段还可以更长；该长度由EA0和EA1属性控制（这里EA表示扩展地址）。如果该属性位被设为0，该标志称为EA0，表示下一个字节包含了地址字段的后续部分；如果该标志是1，该字段被称为EA1，表示地址段的结束。

10位格式的DLCI字段是最常使用的。尽管如此，当使用3字节寻址时，DLCI字段长度是16位。当使用4字节时，长度则变为23位。

帧中继标准使用以下的方法来为用户和网络分配DLCI地址：

• 0——供LMI虚电路使用
• 1-15——保留位
• 16-991——用户用于标识PVC和SVC
• 992-1007——供内部网络连接的传输服务使用

- 1008-1022——保留位
- 1023——用于控制数据链路层

因此，所有的帧中继接口都将976 DLCI地址分配给用户终端设备。

数据字段最多可以包含4 056字节。

C/R字段具有HDLC家族协议的一般含义——“命令——响应”属性。

DE，FECN和BECN字段被用于控制流并支持虚电路的特定QoS。

21.4.2 QoS支持

对于每个虚拟连接，定义了多个参数。每个参数都与数据传输率相关，并影响着QoS。

- **承诺信息率**（commited information rate）（CIR）——确认的网络传输用户数据的确定的信息率。
- **承诺突发大小**（committed burst size）（Bc）——确认的网络根据预先确认的间隔时间T（称为突发时间）需要传输的确定的突发通信量大小。（例如，该用户最大的字节数）
- **过量突发大小**（excess burst size）（Be）——在预定的时间内，网络试图传输的超过已确定B_c值的突发量。

这些参数是单向的，因此一条虚电路需要为每个方向确认不同的$CIR/B_c/B_e$值。

假设之前提到的参数都确定了，那么T根据如下公式定义：$T = B_c/CIR$ (21.1)

也可以先指定CIR和T的值，这样突发量B_c就变成计算得出的值。通常，为了控制突发量，T间隔在传输数据时一般是1～2秒，传输语音时则是几十或者几百毫秒。

图21-6表示了CIR，B_c，B_e和T之间的关系。（R是访问链路的速率，f_1－f_5是帧。）

建立虚连接时用户和和服务提供商之间必须达成约定的主要参数为CIR。对于PVC，这种约定是提供网络服务合同的一部分。当建立SVC时，该约定利用Q.933协议自动生成，所需的其他连接参数——CIR，B_c和B_e——在连接请求分组中传输。

因为数据速率是在特定时间间隔内测量的，T则为检查约定条件所耗费时间的衡量参数。一般说来，在这个间隔内，用户禁止使用高于CIR的平均速率传输数据。

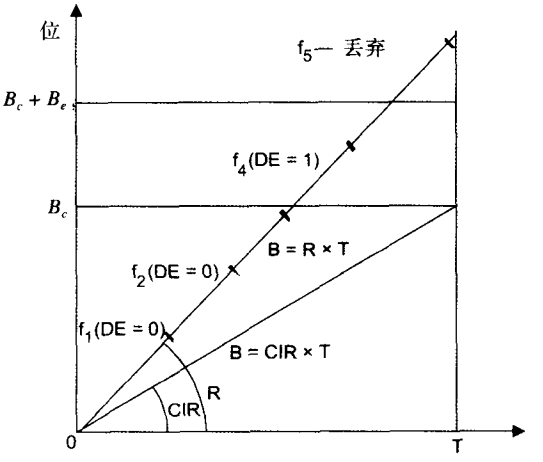


图21-6 网络对用户行为的反应

如果用户不遵守该约定，那么网络将无法保证该类帧的传输，并且会给它们把**丢弃合格**（discard eligibility）（DE）属性设为1。这意味着这类数据将是最先被丢弃的。具有该属性的帧只有在交换机过载的时候才会被延误。如果网络没有超负荷，DE = 1的帧还是会被传递到目的地。

这样的冗余行为仅当用户在整个T阶段通过网络传输的数据总量没有超过$B_c + B_e$才会发生。如果超过了该阀值，该帧会立刻被丢弃，而不设定DE属性。

图21-6展示了在时间段T内传输到网络的5个帧。这段时间内的平均数据率是R b/s，超过了CIR值。帧f_1，f_2和f_3被传入网络因为数据总量没有超过B_c阈值。因此，它们被传输且DE设为0。数据f_4被添加到f_1，f_2，f_3之后，超过了B_c值但是没有超过$B_c + B_e$阈值。于是，f_4也被传输了，然而DE被设为了1。数据f_5被添加到之前传输的帧之后，超过了$B_c + B_e$阈值，因此将它丢弃。

为了控制约定参数，所有的帧中继交换机支持漏桶算法。该算法属于第20章提到的令牌桶算

法同类。它也允许控制平均速率和突发流量，尽管使用不同的方式来实现。

漏桶算法使用计数器C来计算从用户收到的字节数。每隔T秒，该计数器被减去B_c值（或者被设为0如果该值小于B_c）。所有不会使C计数器超过B_c阈值的帧数据都被传输到网络中，DE设为0。所有使C计数器超过B_c阈值，但不超过$B_c + B_e$的帧数据也被传输到网络中，但是这次DE被设为1。最后，所有使计数器超过$B_c + B_e$的帧都被交换机丢弃了。

用户也可以和网络服务供应商签订在特定虚电路上只涉及部分QoS参数的约定。

例如，可以只使用CIR和B_c参数。这种变形提供了更好的服务质量，因为用户帧永远不会被交换机立刻丢弃。交换机仅仅标记超过B_c阈值的帧DE = 1。当网络不拥塞时，该类虚电路上的帧总是被传递给用户，尽管用户持续违反了他与网络所达成的约定。

还有另一种流行的QoS需求，只约定了B_c阈值，而CIR被认为是0。所有的帧都会被标记为DE=1。尽管如此，它们都被发送到网络上，只有在B_e阈值超出时才会被丢弃。 在这里检查时间T被算为B_e/R，R是该链路的访问速率。

从以上描述可以看出，漏桶算法相对于令牌桶算法对网络的流量控制更加严格。令牌桶算法允许流量根据网络低活动时期的突发量然后使用突发阶段的累积总和余量。漏桶算法没有提供这种可能性，因为C在每次T过期时被强制设置为0，不管在该段时间内从用户接收了多少数据。

图21-7显示了有五个远端部门帧中继网的例子。一般说来，对网络的访问由带宽超过CIR值的连接确保。然而，在这个例子里面，用户付费来获得需要的CIR，B_c和B_e值，而不是连接带宽。比如，如果一根T1的线路被用作访问链路，用户要求的服务是CIR = 128 Kb/s，他只会为128 Kb/s付费。T1线路的实际数据率是1 544 Mb/s，只会影响可能突发量——B_c+B_e的上限。

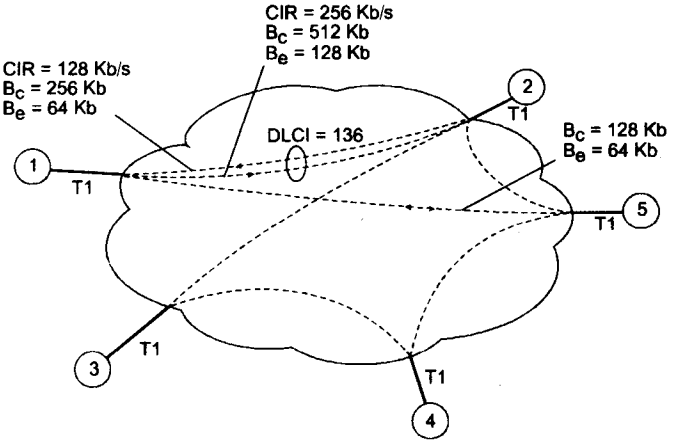


图21-7　利用帧中继网

QoS参数对于虚电路的不同方向可能不同。例如，在图21-7中，用户1通过虚电路DLCI = 136连接到用户2。从用户1到用户2方向，电路的平均数据速率为128Kb/s，B_c为256kb（间隔T使用秒来衡量），B_e为64Kb。当反方向传输数据时，平均数据速率能达到256Kb/s，B_c= 512Kb，B_e= 128Kb。

预留平均带宽和最大突发量是帧中继网中保障QoS的主要机制。

约定一定要确保虚电路平均数据速率总和不能超过网络交换机端口的处理能力。申请永久电路时，管理员主要负责这一点。建立交换电路时，该责任则由交换机软件来代理。当双向的强制规则明确后，网络通过丢弃DE = 1的帧和超过$B_c + B_e$阈值的帧来确保网络正常工作，避免拥塞。

帧中继技术定义了一种可选的流控制算法，同时是一种通知终端用户路由器拥塞的机制（被未处理的帧超载）。转发明确拥塞提醒（FECN）位通知接收端关于这个情况。根据该位的值，接

收端必须通知发送端（使用高层协议例如TCP-IP、SPX之类）减轻帧传输的密度。

回发明确拥塞提醒（BECN）位通知发送端关于网络拥塞，它要求发送端立刻降低传输速率。通常BECN位是被帧中继访问设备来处理的——路由器、多路复用器和CSU/DSU设备。帧中继协议并不要求当设备收到FECN和BECN位被设置的分组时立刻停止帧传输，尽管在X.25网络是这样的。这些数据位是用来为高层协议（TCP，SPX，FTP之类）作为减少传输速率的提醒而用。由于不同协议对接收端和发送端的流量控制是使用不同方法来实现的，所以帧中继开发者考虑了关于网络拥塞控制信息双方向的传输。

21.5 ATM技术

ATM技术是作为一种称为**宽带-ISDN**（broadband-ISDN，B-ISDN）的新一代综合服务的统一传输发展起来的。原则上，ATM为继ISDN失败后第二次组建聚合网络的尝试。相对于帧中继初始针对弹性网络流量的目的，ATM开发者的目的更加广阔和富有野心。

根据开发者的计划，ATM必须具有以下几种能力：

- 单一的传输系统，能够同步传输计算机和对延迟非常敏感的多媒体（语音和视频）流量；对于每种流量的QoS要符合各自对应的要求
- 传输率层次从每秒几十兆字节到每秒几千兆字节，并且有足够带宽留给关键应用
- 能够适用现存的物理连接或者物理协议架构：PDH，SDH或者高速局域网
- 和现有的局域网和广域网相互作用，例如IP，SNA，以太网和ISDN

很有必要指出，大部分目标都成功完成了。从20世纪90年代中期开始，ATM变成一种实际可用的技术，能够确保为网络用户支持最完整稳定的QoS参数。不仅如此，和其他虚电路技术类似，ATM在解决流量工程问题方面提供了广阔的空间。

ATM标准的开发（RFC 2514、RFC2515、RFC2761、RFC3116等）的制定由很多参与ATM论坛电讯设备制造商和通信供应商执行。一些特别的ITU-T和ANSI委员会也参与了这项工作。

除了运行在大型通信供应商主干网上ATM技术显著的成功，它也表现出了局限性。因此，ATM没能够把其他的技术挤出舞台，变成唯一的电信网络传输技术，即便这一点在20世纪90年代中期由于ATM技术显著的优势，看上去是不可避免的。理论上，ATM能直接被应用层协议使用，网络可以不依靠IP和TCP/UDP直接运行。ATM提供了满足该需求的许多特性，包括支持各类信息流，可扩展性，和一个原生的、复杂的路由协议。尽管如此，这一点只有在网络变成技术上同质时才会实现。为了这个目的，所有的服务供应商的所有网络都必须支持ATM。很显然，这样的方式和互联网的主要目的相悖，因为互联网的目的就是为了支持自身的传输技术和网络层来把这些关键网络合成一个唯一的互联网。

于是，IP在20世纪90年代中期之后取得了网络层的统御地位。对于ATM，它成为一种很多不同网络互联运行的技术基础。ATM和IP互联的问题将会在第22章提到。

21.5.1 ATM运行的主要原理

ATM网络有大规模广域网的典型特征。工作站（终端节点）通过独立的链路连到低层交换机。这些交换机依次连接到高层交换机。ATM交换机使用基于虚电路技术的20字节终端节点地址对流量进行路由。对于专有ATM网络，定义了专用NNI（PNNI）路由协议，路由器可以根据它自动生成路由表，而且符合流量工程学的要求。通常，基于E.164标准的地址在公用ATM网络中使用，简化了这些网络与电话网络的互联。ATM地址有一个类似电话号码的和IP地址的架构层次。这确保ATM可扩展到任何需求层，甚至到世界范围的网络。

为了加速大规模网络的交换，虚拟路径的概念出现了。虚拟路径连接在ATM网络中具有公用

路由的虚电路，该ATM网络连接源节点和终端节点；或者连接两个路由器间共享部分路由的虚电路。该属性提高了ATM的延拓性，因为它大量减少了主干路由器需要支持的虚拟连接数，由此提高了操作的效率。

ATM标准没有引入网络层实施的规范。在这方面，它基于SDH/SONET技术并采用了速度层次。因此，对网络用户提供的起始访问速率是STM-1速率，即155 Mb/s。ATM主干设备以更高的速度运行——STM-4 622Mb/s和STM-16 2.5Gb/s。也有ATM设备支持PDH速度的，比如34Mb/s和45Mb/s。

所有之前列举的ATM技术特性还不足以说明它是一种特别的技术。其实，它是一种基于虚电路的广域网技术，ATM技术的特点在于对不同网络流的质量服务。

把ATM和其他技术区别开来的，是针对所有网络流的整合QoS支持。

为了实现该特性，ATM开发者仔细分析了所有网络流的特点，并为它们分类。你们应该在第7章中考虑对于不同应用的QoS时了解到这个分类。ATM分类把所有的数据流分为五类：A、B、C、D和X，前四种类别代表了典型的应用。这些应用对于网路分组延迟和丢失有一个稳定的需求集。该类另外一个应用显著的特点是它们产生的数据流具有稳定波特率（SBR）和可变比特率（VBR）特征。类别X是预留给特殊应用程序的，它们的特征集合和QoS需求不能被纳入前四类。

尽管如此，任何分类方法中的数据流分类个数都不意味着对问题本质上的解决办法——具体说，就是找到在同一通道内成功支持所有弹性和延迟敏感数据流的方法。这些分类的需求基本上都是互相排斥的，一个典型的例子就是对于帧大小的要求。

弹性数据流受益于可以扩展帧的大小，因为这减少了控制信息的额外开销。在以太网的例子中，当把数据字段的大小从最小的46字节提高到最大值1 500字节时，你可以发现有效数据率能够提高两倍以上。当然，帧不能无限扩大因为这样分组交换就会失去意义。然而，对于同速率水平的弹性数据流，几千字节大小的网络分组是可以接受的。

延迟敏感数据流，正好相反，当帧大小是几十字节时工作得更好。当使用大的帧时，会出现下列两种不好的影响：

- 在队列中低优先级帧的较大延迟
- 分组延迟

我们以话音数据流为示例来考虑这些影响。

我们知道延迟敏感数据流的队列中等待时间（*waiting time in queue*）能够通过使用优先级队列来减少。然而，如果分组大小能在很大范围内波动（例如，从29到4 500字节，如同在FDDI技术中那样），那么即使话音分组在交换机被授予最高优先级，计算机分组的等待时间可能会长得无法接受。举个例子，一个4 500字节的网络分组会以2Mb/s的速率（帧中继访问端口的典型操作速率）在18毫秒内传到输出端口。当合并数据流时，有必要通过同一端口同时传输144个语音采样。我们不想中断分组传输，因为在一个分布网络内，通知相邻交换机该中断的信息和相应的恢复通讯所带来的额外开销高的无法接受。

封包延迟（packetizing delay）是第一个话音采样等待整个网络分组完全形成然后被发送到网络内的时间。该延迟的机制如图21-8所示。

编码器按照统一的间隔产生话音采样。例如，图中所示的PCM编码按照8KHz的频率采样——就是说，每隔125μsec。如果使用以太网中最大的帧来传输话音，每个帧将会携带1500个话音采样，因为每个采样被编码为1字节大小。结果就是，第一个进入以太网帧的采样必须等待$(1500-1)\times 125 = 187375\mu sec$，大约是187毫秒。这个延迟对于话音来说非常显著。根据ITU-T建议，该延迟不应该超过150毫秒。这里很需要注意的一点是，帧延迟和协议的比特率没有关系。它只取决于编解码器操作频率和帧的数据字段大小。这一点区别于随着比特率提高而减少的队列

延迟。

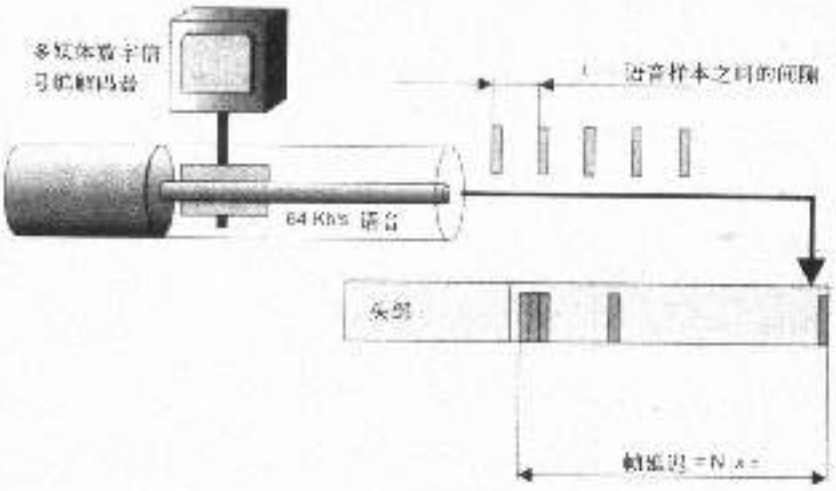


图21-8 分组延迟

ATM帧的数据字段48字节大小是针对突发数据流和延迟敏感数据流的折中方案。换句话说，可以认为这个折中是被电话专家和计算机专家实现的——前者坚持数据字段大小应该是32字节，后者要求64字节。因为ATM域的这种小而固定的大小，它后来被称为**单元**（cell）。

对于一个48字节的数据字段，单独的ATM信元通常每隔125μsec携带产生的48个语音采样。因此，第一个采样在整个单元被发送到网络前需要等待大概6毫秒。这正是电话专家要求缩小帧大小的原因。6毫秒已经很靠近语音传输质量下降的阈值。如果单元大小选为32字节，分组延迟大概是4毫秒，能够很好的保证语音传输质量。而计算机专家要求加大数据字段到64字节也很有道理，因为如果能够满足该条件，有效数据率能够被提升。当使用48字节数据字段时，多余服务数据是10%，当使用32字节数据字段时，会被提高到16%。

对于包含53字节的分组，以155Mb/s速率传输到输出端口需要少于3μsec时间。因此，该延迟对于分组每隔125μsec传输的数据流是可以接受的。

为了保证分组含有目标地址信息，并且确保控制信息比例不超过数据字段大小，ATM技术实施了广域网的一个标准，叫做基于虚电路技术的单元传输。整个虚电路编号长度是24字节。足够支持基于ATM技术的大规模广域网上，交换机每个端口上的大量虚拟连接。

很有必要指出的是，在ATM中使用小单元为高质量的延迟敏感数据流提供了理想的条件。相应的代价就是当交换机运行在高速率下的高负荷。不要忘记基于任何技术的路由器上的工作量都直接与每个时间单元处理的分组或者帧数量挂钩。显然，处理数据字段48字节长度的单元，相对于在以太网内操作1500字节域的交换机，给ATM交换机带来的负荷高很多。因为这个原因，ATM交换机无法长时间以超过接口速率限制622Mb/s工作。最近，它们才开始支持2.5Gb/s。

选择小的固定大小信元本身并不能解决在同一网络中合并不同类型数据流的问题。实际上，它仅仅是保证找到了解决方法的前提。为了完全解决这个问题，ATM技术发展和实现了按需提供带宽和*QoS*控制的概念，该概念在帧中继技术中加以实施。

在ATM技术中，每种数据流类别都有一组定量的参数必须被应用程序指明。对于A类数据流，必须指出应用程序发送数据到网络中的稳定速率；对于B类数据流，必须指明最大可能速率，平均速率，和最大可能突发量。对于语音数据流，还可以指定发送端和接收端之间同步的重要性，也可以通过指定延迟和延迟变量的上限值来提供定量的特征。

ATM支持以下的虚电路数据流的主要定量参数：

- **峰值信元率**（peak cell rate）（PCR）——数据传输的峰值速率
- **持续信元率**（sustained cell rate）（SCR）——数据传输的平均速率
- **最小信元率**（minimum cell rate）（MCR）——数据传输的最小速率
- **最大突发大小**（maximum burst size）（MBS）——数据突发的最大值
- **信元丢失率**（cell loss ratio）（CLR）——丢失信元的比例
- **信元传输延迟**（cell transfer delay）（CTD）——信元传输的延迟
- **信元延迟变量**（cell delay variation）（CDV）——信元传输延迟的变化范围

速率参数使用每秒信元数来衡量，MBS使用信元数来衡量，时间参数用秒来衡量。MBS在提供了平均速率的情况下，说明应用程序在PCR时能传输的信元数量。丢失单元比例是丢失的单元数量占通过该虚拟连接传输的总单元数的比例。因为虚拟连接是双向的，所以可以分别指派每个方向的数据传输值。

ATM技术以一种非常规的方法来诠释QoS术语。通常，QoS的特征是带宽参数（在这里，就是RCR、SCR、MCR和MBS）、分组延迟参数（CTD和CDV）和分组传输可靠性参数（CLR）。在ATM中，信息率特征被称为流量参数。它们没有被纳入QoS参数，尽管原则上它们是的。在ATM中，QoS参数仅包括CTD、CDV和CLR。网络尽力确保服务水平，这些服务为流量参数、信元延迟和丢失信元比例提供所需取值。

应用程序和ATM网络之间的约定被称为流量协议。和帧中继网中协议最大的区别是，除了带宽参数之外，可以为某一种数据流类别选择信元延迟参数和信元传输可靠性参数。在帧中继中，只有一种数据流类别，它的特征只有带宽参数。

如果对于应用程序来说，带宽参数和QoS参数的支持无关紧要，那么就可以通过在建立连接的请求中指定最佳属性来忽略它们。这种类型的数据流也被称为**未指明波特率**（unspecified bit rate，UBR）数据流。

当总结了和特定虚拟连接相关的流量协议后，ATM网络提供了几种确保QoS需求的协议和服务。对于UBR流量，网络尽可能地分配资源（例如，当前可分配资源），既然它们并没有在申请了特定QoS参数的虚拟连接中使用。

21.5.2 ATM协议栈

ATM协议栈如图21-9所示，协议在终端节点间分配情况和ATM交换机见图21-10。

ATM协议栈对应于ISO/OSI模型的低层，它包含ATM适配层（AAL）、ATM层和物理层。在ATM和OSI协议层之间，没有直接对应关系。

网络中的高层

<table>
<tr><td rowspan="2">ATM适配层
ATM (AAL1-5)</td><td rowspan="2">聚合子层（CS）</td><td>聚合子层的公共部分</td></tr>
<tr><td>聚合子层特殊服务部分</td></tr>
<tr><td colspan="2">分割和重编译子层（SAR）</td></tr>
<tr><td colspan="3">ATM适配层
（分组路由，多路复用，流控制，优先级处理）</td></tr>
<tr><td rowspan="2">物理层</td><td colspan="2">传输调合子层</td></tr>
<tr><td colspan="2">媒质相关子层</td></tr>
</table>

图21-9　ATM协议栈的架构

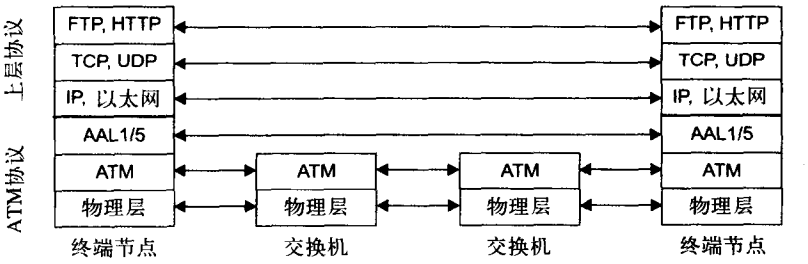


图21-10 ATM协议在交换机和终端节点间的分布

21.5.3 ATM适配层

AAL将ATM网络高层协议消息转换为ATM指定格式单元的AAL1-AAL5协议集合。这些层的功能可以很方便地对应到OSI运输层（例如TCP和UDP）。AAL协议仅在网络终端节点操作，就像大多数网络技术的传输协议一样。

AAL层的每个协议处理特定类别的用户数据流。在标准化的开始阶段，每种数据流类别有自己对应的AAL协议，该协议在终端节点收到高层协议的网络分组，为应用程序需要的特定虚拟连接使用合适的协议来请求数据流和QoS参数。随着ATM标准不断进化，这种数据流类别和AAL协议间清晰的对应消失了。现在可以对同一种数据流类别使用不同的AAL协议。

AAL包含两个子层。

- 低级层是**分割和重编译（SAR）层**。这一部分不依赖于AAL协议类型（或者，正在传输的数据流类别）。它对AAL从高层协议收到的消息进行分段。创建ATM信元后，SAR提供给它们合适的头部然后传入网络。
- 高层部分的AAL是**会聚子层（CS）**（convergence sublayer，CS）。该子层依赖于正在传输的数据流类别。CS协议解决诸如保障传输节点和接收节点同步（对于需要这样同步的数据流而言）、控制和恢复用户数据中的错误位、传输中计算机协议分组的完整性控制（X.25或者帧中继）之类的问题。

为了完成任务，AAL协议使用位于AAL头部的控制信息。收到通过虚电路传来的信元后，SAR使用AAL头部收集源消息（通常分为多个ATM信元）。AAL头部对于ATM交换机是不可见的，因为它们位于信元内48位的数据字段中，而高层协议对于这部分不敏感。收集源消息后，AAL检查AAL帧的头和尾来获取控制字段，然后根据检查结果来决定收到的信息是否正确。

传输用户数据时，AAL协议并不会恢复丢失的或者受损的数据，它能做的是通知终端节点该事件。假设数据丢失或者损毁是很少见的，这样做的目的是提高ATM交换机的操作。恢复丢失数据（或者忽略这样的事件）被代理给ATM协议栈外的更高层协议。

AAL1协议通常为A类流量提供CBR,一般来说是给如数字视频或者语音这样延迟敏感的数据流。ATM模拟一般的相用数据链路来传输这样的数据流。AAL1头部在ATM信元中占用数据字段的1～2字节，相应的留出47或者46个字节给用户数据。头部的一个字节是分配给信元编号的，这样接收节点能够判断是否收到了所有的信元。当发送语音数据流时，对于每个采样的时间戳是已知的，因为它们之间间隔为125μsec。因此，如果这个信元丢失了，通过简单将下一个信元位移125*46μsec来更正是可行的。丢失语音采样的几个字节并不要紧，因为在接收节点的设备能够平缓这些信号。AAL1协议的任务包括平滑处理目标节点收到的不均匀信元。

AAL2协议为了传输B类流量而开发，后来它被排除在ATM协议栈之外。现在，使用AAL1、AAL3/4或者AAL5协议来传输B类数据流。

AAL3/4协议处理局域网内的突发数据流类型。这是一个VBR，并且数据流使用一种防止信元

丢失的方式来传输。尽管如此，信元有可能被交换机延误。当传输信元时，AAL3/4协议使用一种复杂的差错控制过程。为了实现这个目标，它对源信息的每个部分进行编号，并为每个信元提供校验。然而，如果信元丢失或者受损，该层并不试图恢复它们。相反，它完全丢弃整个消息——就是所有其他剩下的信元。这是因为对于计算机数据流或者压缩语音来说，即使丢失一个信元也是很严重的错误。AAL3/4协议合并了AAL3和AAL4协议的结果，它们都能够使用面向连接和无连接的协议来确保对计算机数据流的支持。正是因为它们使用类似的头部格式和操作逻辑，AAL3和AAL4协议才能够被合并。

AAL5协议是AAL4协议的简化版本。它操作起来更快，因为不需要针对每一个消息信元计算校验和。不同的是，它为整个消息计算校验和，并将它放到消息的最后一个信元中。起初，AAL5协议是为了传输帧中继网的帧而开发的。然而现在，它用来传输所有类型的计算机数据流（RFC 2684）。除了和发送端与接收端之间同步相关的参数之外，AAL5还能够支持其他QoS参数。因此，它被用来广泛支持和计算机数据流传输相关的所有类型流量（例如， C类和D类流量）。一些设备制造商用AAL5来服务CBR流量；他们把同步任务交给高层协议代理。AAL5不仅在终端节点运行，也在ATM交换机上运行。然而，在交换机上它只操作和用户数据传输无关的控制功能并且支持牵涉到建立虚拟连接的高层服务协议。

在AAL层和应用程序之间，有一个特殊的接口需要使用ATM网络传输数据流。使用该接口，应用程序（例如，一个计算机网络协议或者一个语音采样模块）通过判断数据流类型和它的QoS参数来请求所需的服务。ATM技术允许两个定义QoS参数的方法：直接由各程序定义和根据数据流类型使用默认值。后面一种方式简化了应用程序开发者的任务，因为它将信元传递的最大延迟和延迟变量的决定权交给了网络管理员。

AAL协议本身并不能确保所需的流量和QoS参数。为了遵循流量约定的条件，需要保证整个虚电路上所有的交换机间协调操作。这个任务由ATM协议完成，它保障不同虚拟连接的信元在所需QoS级别上传输。

21.5.4 ATM协议

ATM协议在ATM协议栈中的地位与IP在TCP/IP协议栈或者LAP-F在帧中继协议栈中的地位差不多。在虚电路建立和配置后，ATM协议使用交换机传输信元，这意味着此项任务是基于端口交换表的。

ATM协议通过使用**虚电路序号**（number of the virtual circuit）来实现交换，这在ATM技术中分为两部分：

- **虚拟路径标识符**（VPI）（virtual path identifier，VPI）
- **虚拟通道标识符**（VCI）（virtual channel identifier，VCI）

除了这个主要任务，ATM协议还执行诸如允许网络用户监控数据流协议、标记冲突信元、在网络拥塞时丢弃冲突信元以及控制信元流来提高网络性能（当然，在所有虚电路都监控数据流协议的前提下）等功能。

ATM信元格式如图21-11所示。

通用流控制（*generic flow control*）字段不仅仅用于终端节点和第一个网络交换机间的交互过程，其具体的功能尚未定义。

*VPI*和*VCI*字段分别占用1和2个字节。这些字段指明被拆分为最高位部分（VPI）和最低位部分（VCI）的虚拟连接号。

有效负荷类别标识符（*payload type identifier*）包含3位。它指明单元携带的数据类型——用户数据或者控制数据（例如，当建立一个虚拟连接时）。并且，该字段其中的一位用来显示网络拥塞。

该位被称为显式堵塞转发标识符（EFCI），和帧中继技术中的FECN有着同样的功能，它将网络拥塞的信息按照数据流动的方向传递。

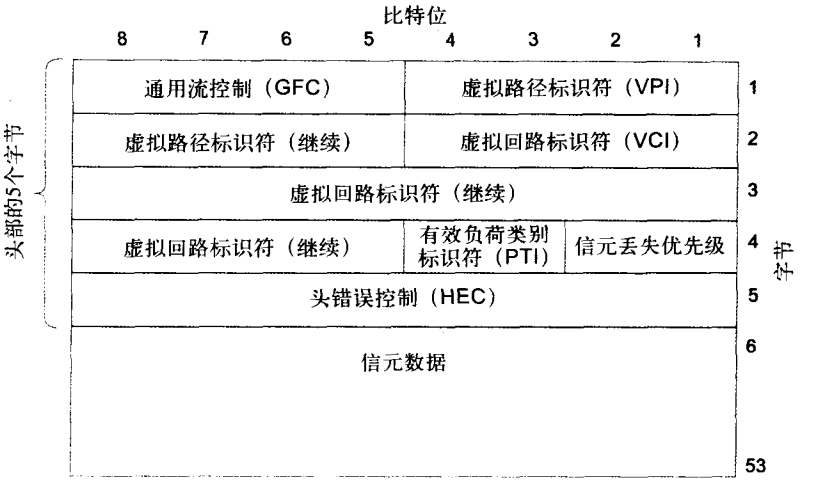


图21-11 ATM信元格式

该技术中的*信元丢失优先级*（*cell loss priority*，*CLP*）和帧中继技术中的DE字段起同样的作用。ATM交换机用这个字段来标记违反了该连接QoS参数协议的信元。如果网络拥塞发生，这样的信元将会被丢弃。因此，CLP = 0的信元对于网络来说是高优先级的，CLP=1则是低优先级。

头部错误控制（*header error control*，*HEC*）字段包含为信元头部计算的校验和。该校验和通过Hamming纠错码计算得出。因此，它不仅能够发现错误，还能够纠正所有的违规错误（并且，在合适的情况下，甚至一些重复错误）。HEC字段不仅确保头部中的错误检测和纠正，还能够发现ATM技术的物理层SDH字节流的边界，或者在基于信元的物理层的位元流中发现边界。在ATM中，不像SONET/SDH技术中的STS-n/STM-n帧那样具有能够监测ATM信元边界的指针（类似于在T1/E1通道中的虚拟载体的边界判定指针）。于是，ATM交换机为位于STM-n帧中数据字段的5字节序列计算校验和，如果校验和显示ATM信元头部是正确的，那么第一个字节变成信元边界。否则，偏移1个字节，然后继续这个操作。也就是说，ATM技术检查同步SDH帧中的异步ATM信元流，或者检查基于信元的物理层位流。

ATM交换机可以运行于两种模式：

- **虚拟路径交换**（virtual path switching）。在该模式下，交换机仅仅基于VPC字段来转发信元，忽略VCI字段。通常，大规模区域主干交换机用这种方式运行。它们仅依靠对应于地址聚合概念的虚电路序号的最高位在网络之间传递信元。因此，一个虚拟路径对应于整个一系列虚电路聚合。
- **虚电路交换**（virtual circuit switching）。传递信元到本地ATM网络后，本地的交换机根据VPI和VCI来交换信元。然而，为了交换它们仅需要虚拟连接序号的最低位；于是，它们仅基于VCI工作，VPI保留不变。

要创建SVC，ATM使用图21-11中未显示的协议。在这里，这个方式类似于ISDN的方法。为了建立一个连接，发展一个独立的协议，Q.2931，它可以归为一个网络层协议。该协议和Q.931及Q.933有很多相似之处（包括编号）。然而很显然，为了支持相关的几种数据流类别和额外的QoS参数，也有一些更改。Q.2931协议依赖于能够保障Q.2931帧稳定传输的一种被称为SSCOP的复杂的

数据链路层协议。当SSCOP帧传递到目标交换机时，SSCOP也运行在AAL5协议上，后者是将SSCOP帧拆分为ATM帧以及将信元收集为帧所必须用到的。

注释　继引入了UNI3.1接口之后，Q.2931出现在ATM协议栈中。UNI3.0使用的是Q.93B协议。由于Q.2931和Q.93B之间不兼容，UNI3.0和UNI3.1接口版本也不兼容。UNI4.0版本提供了对UNI3.1的向下兼容，因为它和UNI3.1是基于相同的控制协议。

使用Q.2931创建的虚拟连接既可以是单工的（单向），也可以是双工的（双向）。

Q.2931协议也允许建立点对点和一对多的虚拟连接。第一种情况为所有基于虚电路的技术所支持，第二种方式仅对ATM可用，是多播的一种模拟，除了单一多播节点之外。建立一对多连接时，引导节点是连接的发起者。首先，该节点与另一个节点建立一个虚拟连接，然后，使用一种特殊的调用，它添加新成员到这个连接。发起者成为连接树的根，其他所有参与节点变成“叶子”的角色。引导节点发送的消息被所有连接的叶子收到；然而，某一个特定叶子发送的消息（在一个双工连接中）仅仅被引导节点接收。

Q.2931协议的网络分组试图建立的SVC和Q.933协议的网络分组有同样的名字和目的，在本章描述帧中继技术的部分有所介绍。然而，它们字段的结构是不同的。

终端节点地址在ATM交换机中是20字节地址。

在公共网络操作时，使用了对应于E.164标准的地址。这种地址格式可变，能被分为很多部分来保障网络间和子网间的分层路由。它比IPv4支持更多的层次等级，在这方面和IPv6类似。

地址的最后6字节分配给**终端系统标识符**（end system identifier）（ESI）。该字段和ATM节点的MAC地址作用相同。它的格式也对应于MAC地址格式。

ESI地址是由设备制造商根据IEEE规则赋予终端节点的。这意味着头3位字节包含制造商代码，后3位是制造商必须保证唯一的序列号。

在一个私有ATM网络工作时，地址格式通常对应于略微改动后的E.164格式。

当终端节点连接到ATM网络时，它会运行一个所谓的注册过程。终端节点发送它的ESI地址给交换机，交换机向终端节点发送最高位的地址部分——也就是节点所连接的网络编号。

除了地址部分外，终端节点使用Q.2931协议的呼叫建立网络分组来请求建立一个虚拟连接，包含描述数据流参数和QoS要求的部分。当这样的分组到达路由器时，路由器必须分析这些参数然后决定它是否有足够的资源来服务一个新的连接。如果有，那么会接受一个新连接，路由器根据目标地址和路由表来发送呼叫建立网络分组。如果可用资源不足，那么请求会被拒绝。

21.5.5　ATM协议服务和流量控制的种类

为了支持不同虚拟连接要求的QoS，并且保证网络资源的合理使用，ATM网络实施多种与用户数据流相关的ATM层次的服务。这些是ATM网络的内部服务。它们试图使用AAL协议来支持不同的用户数据流类别。然而，相对于操作在终端节点的AAL协议，这些服务分布在所有的网络交换机上。它们按照终端节点AAL层输入数据流类别来对服务进行分类。ATM层的服务是终端节点建立连接时，通过UNI使用Q.2931协议请求得到的。当请求这些服务时，如果有到AAL的请求，必须指明它的类别、数据流和QoS参数。这些参数采自AAL层类似的参数，或者来自根据服务类别定义的默认值。

ATM协议层有5个服务类别，它们以同样名字的服务来支持：

- **CBR（稳定比特率）**（constant bit rate，CBR）——针对CBR数据流的服务
- **rtVBR（实时可变比特率）**（real-time Variable bit rate，rtVBR）——针对VBR数据流的服务，需求稳定的平均数据速率并且发送端与接收端间要求同步

- nrtVBR（非实时可变比特率）（non real-time Variable Bit Rate，nrtVBR）——针对VBR数据流的服务，需求监测稳定的平均数据速率，但不要求发送端与接收端间同步
- ABR（可用比特率）（available bit rate，ABR）——针对VBR数据流的服务，需求最小数据率，不要求发送端与接收端间同步
- UBR（未指明比特率）（unspecified bit rate，UBR）——针对数据流的服务，数据传输不提供任何参数，也不要求发送端与接收端间同步

大部分服务类别的名字都和它们服务的用户数据流类别的名字相一致。然而，必须理解ATM层次服务是ATM网络内部服务，通过AAL层实现隐藏于应用。

*CBR*服务（*services*）适用于支持同步应用数据流——语音，专用数字线路仿真等。当一个应用建立CBR类型的连接时，它请求信元峰值速率（PCR）即在信元不丢失的情况下，连接支持的最大速率。应用也会请求如下QoS参数：CTD、CDV和CLR。

于是，数据按照预先请求的速率，不会超过，同时大多数情况下也不会小于该速率的方式通过该连接传输，尽管速率降低是可能发生的——例如，使用CBR服务类别传输压缩语音时。所有通过该站点以更高速率传输的信元，都被第一个网络路由器控制并且标识为CLP=1。当网络拥塞发生时，这样的信元就会被网络丢弃。其他延迟的、不符合CDV间隔约定的信元也被认为是低优先级的，被标为CLP=1。

对于CBR连接，没有类似在T1/E1线路中，速率必须是64Kb/s倍数这样某种不连续PCR的限制。

相比于CBR服务，当应用程序和网络间请求建立连接时，*VBR*服务要求更复杂的过程。除了PCR，VBR应用还要求两个额外的参数——SCR，应用允许的平均数据传输速率，和MBS。MBS通过ATM信元数来衡量。用户只能短时间超过PCR阈值允许的速率，而在此期间数据量不能超过传输的MBS。这个时间段被称为**突发容忍（BT）**（burst tolerance）。网络根据三个指定参数计算该值：PCR、SCR和MBS。

如果PCR以超过BT的时间运行，信元就会被设置为违反服务协议——CLP属性被置为1。

*rtVBR*服务设置了与CBR服务同样的QoS参数，nrtVBR服务则仅仅支持数据流参数。对于两种VBR数据流类型，网络支持一定程度的CLR，而该CLR要么是建立连接时指定的，要么是针对某种数据流类别设定的默认值。

对于控制数据流和QoS参数，ATM使用一种能检查用户是否遵守相关参数如PCR、CDV、SCR、BT、CTD约定的算法，即所谓的*通用信元率算法*。它模仿帧中继技术中使用的更改后的漏桶算法运行。

对于很多突发性强的数据流的应用，有必要预测哪些在建立连接时就需要确定的数据流参数。例如，两个通信局域网之间的事务处理和数据流就是无法预测的。数据流密度的变化如此之大，以至于不可能总结出该网络任何合理的协议。

相对于CBR和两种VBR服务，UBR服务并不支持数据流参数和QoS参数。UBR只提供尽力传递，没有任何保障。特别是当超过网络带宽限制时，UBR服务对于无法获取任何数据流参数的、不可预测性的突发应用提供了部分解决方案。

UBR服务主要的缺点是缺乏流量控制机制和无法顾全其他类型的数据流。UBR连接即使在网络拥塞时，也会继续传输数据。网络交换机能够缓冲一部分传来的信元，但是最终缓冲溢出后，信元会丢失。由于UBR连接完全没有约定的数据流或者QoS参数，它们的信元将会首先被丢弃。

*ABR*服务类似于UBR，提供了超出带宽的可能性。但是因为数据流控制技术，它能够保证可靠的传输，提供了一些信元传递的保障。

ABR是第一种能够保障突发数据流稳定可靠的ATM层服务，因为它能够找出在一般网络数据流中，别的服务类别不需要的时间片并且填充上自己的信元。

类似于CBR和VBR服务，当建立ABR类的连接时，需要确定PCR参数。然而，对于信元传输参数或者突发参数，并没有任何既定协议。于是，网络和终端节点约定出所需最小的传输速率，即MCR。对于终端节点运行的应用程序，这样能够保障小带宽，通常是该应用程序正常运行的最小要求。当约定该协议时，终端节点同意不会超过PCR速率传输数据，而网络则保证保留MCR的带宽。

如果建立ABR连接时，不确定最大和最小速率，那么就默认连接PCR和终端节点访问网络的线路速率相同，MCR默认为0。

ABR数据流针对丢失信元和可用带宽获得了有保障的QoS。对于信元传输延迟，尽管网络尽最大努力来减少它，还是没有办法提供保障。相应的，ABR服务并不适合传输对于传输延迟高度敏感的实时应用数据流。

当传输CBR、VBR和UBR数据流时，缺少对网络拥塞的专门控制。取而代之的是丢弃违反约定信元的机制。同时，使用CBR和VBR的节点尽最大努力不违反数据流协议。如果它们这样做，它们会面临着丢失信元的风险。因此，即使有可用的额外带宽，它们也不会使用。

ABR服务允许使用保留带宽，因为它使用反馈机制来通知终端节点有保留带宽的存在。同样的机制可以帮助ABR服务减少当网络拥塞发生时，终端节点传入网络的数据量。

使用的ABR服务的节点必须定期和数据信元一起发送**特定资源管理（RM）信元**（special resource management cell）。和数据流一起发送的RM信元是**转发资源管理（FRM）信元**（forward resource management），相反方向传递的信元是**反向资源管理（BRM）信元**（backward resource management）。

有若干种**反馈环路**（feedback loop），其中最简单的存在于两个终端站点之间。当它存在时，网络交换机通过ATM协议携带的数据信元中直接堵塞控制字段中的一个特别标志（EFCI标志）来通知终端站点。于是，终端站点通过网络发送一个特别的消息，该消息包含在一个特定BRM信元中，通知源站点需要降低向网络发送信元的速率。

当使用这种方式时，终端站点对于流控制起主控作用，路由器对于反馈是被动的，仅仅通知发送站点关于堵塞的信息。

这样简单的方式有几个弊端。终端站点无法从BRM消息中获得信息，无法得知需要降低发送到网络的数据速率。因此，它只有把速率降到最低，即MCR，尽管可能这样剧烈的降低完全是没有必要的。另外，在一个长距离的网络中，交换机必须在整个期间持续数据缓冲直到拥塞的通知穿过整个网络。注意对于广域网，这个过程可能会很长，因而可能会发生缓冲溢出，这样便无法达到预期的效果。

还有其他的、更加复杂的流量控制机理。在这些方法中，交换机扮演了一个更加主动的角色，发送节点能够获得更多所支持的数据率信息。

在第一种方法中，源节点发送一个指明所能支持的数据率的FRM信元。该消息在虚电路上所经过的每个交换机都能将请求速率改变到根据当前可用资源能够支持的速率。同样也可以保留请求速率，完全不更改它。目标节点收到FRM信元后，把它转换为BRM信元然后反向传输。注意，它也可以减少请求速率。接收到BRM信元中的应答后，源节点就能确切知道可用数据率是多少。

在第二种方法中，每个网络交换机能够即作为源节点，又作为目标节点运作。作为源节点，它能够产生FRM信元然后经由虚电路发送。作为目标节点，它能够根据收到的FRM信元内容反向发送BRM信元。这种方式更快，更适用于长距离网络。

从描述可以看出，ABR服务趋向于对特定虚拟连接支持QoS参数。它也很适合在网络用户中合理分配资源，而该举措，长期看来，将会提高对所有网络用户的服务质量。

ATM使用不同的机制来支持所需的QoS。基于ITU-T和ATM论坛标准，总结出数据流协议和QoS参数，违反这些约定的信元会被丢弃。除开在它们中描述的机制外，实际上所有的ATM设备制造商都根据几种不同优先级实施了信元队列。

服务数据流基于优先级的策略反过来建立在每个虚拟连接的服务类别的基础上。在引入ABR需求之前，大部分ATM交换机实施一种简单的单层服务设计，该设计把VBR数据流设为最高优先级，第二优先级给VBR，第三给UBR。根据这样的设计，CBR和VBR联合有冻结其他类型服务的潜在可能性。这样的设计对于ABR数据流不会正常工作，因为它无法保障信元传输最小速率的要求。要保障这样的要求，必须分配一定的可保障带宽。

为了支持ABR服务，ATM交换机必须实施能够满足服务于CBR、VBR和ABR要求的两层设计。根据该设计，交换机为每个服务类型提高一定比例的带宽。CBR数据流占用用于支持PCR的带宽，VBR数据流占用用于支持SCR的带宽，ABR数据流占用能够保障MCR的带宽。这样就确保了每种连接都能无信元丢失的运行，也不会牺牲CBR或者VBR数据流来传递ABR信元。在该算法的第二层，CBR和VBR数据流需要时能够占用所有的剩余带宽。因为ABR连接已经得到有保障的最小带宽。

为了支持之前提到的服务的正确运作并为所有数据流类别确保特定QoS等级，需要解决一个独立的任务，即使用数据流工程方法来优化ATM网络操作。在ATM中（也可以在帧中继中）使用虚电路技术为解决数据流工程问题提供了很好的前提条件。然而，目前还没有能够实现自动平衡网络负载，动态为虚电路选择路由器的自动化流程。所有和优化路由器相关的工作都需要预先执行，比如使用一些第三方网络优化技术或者模拟软件。在这之后，必须根据选好的路由器手工配置建立PVC。

在ATM网络中，虚电路和虚拟路径路由的选择能够使用PNNI路由协议，该协议不仅考虑到连接的最小带宽，也保留有足够带宽给新建立的连接。

小结

- 虚电路技术包含对路由和网络分组交换的分离操作。这样的网络的第一个网络分组包含有被呼叫用户的地址，并通过配置传递交换机来创建一个虚拟路径。所有其他在交换模式通过该虚电路的网络分组都是根据虚电路编号的。
- 虚电路技术的优点是快速的网络分组交换，这是基于虚电路编号和网络分组地址字段的减少，以及相应的头冗余的降低。缺点包括不能通过两个不同的路由器实现两个用户间的并行数据流和为短期数据流建立虚拟路径的低效率。
- X.25网络是最古老的、也是最饱经测试的WAN技术。X.25的三层协议栈在不稳定和充斥噪音的通信链路上，通过在数据链路层和网络分组层的纠错能力以及数据流控制能力，能够很好地运行。
- 大部分早期的帧中继网络仅支持永久虚电路（PVC）。交换虚电路（SVC）仅在最近时期才在实际中被部署和投入应用。
- 帧中继网络用于传输突发计算机数据流。因此，在带宽保留的过程中，需要指明数据流平均速率、CIR和协调突发大小，B_c。
- ATM技术是对于预先保留虚电路带宽概念的进一步发展，该概念在帧中继中首先实现。
- ATM支持针对用户不同类型的主要数据流类别。包括CBR数据流、典型的电话网络和视频广播网络中的数据流和VBR数据流、典型的计算机网络数据流和传输中压缩的语音或视频。

- 对于每种数据流，用户都可以向网络要求多种QoS参数，包括最大比特率（PCR）、平均比特率（SCR）、最大突发大小（MBS）和多种有关发送端和接收端之间时间控制的参数，这对于延迟敏感数据流是很重要的。
- ATM技术本身没有定义新的物理层。相反，它使用现有的物理层。ATM技术主要的标准是SONET/SDH和PDH的物理层。

复习题

1. 什么参数能被用于描述一个虚拟连接？
2. 如果虚电路经过路径上的物理链路失效，那么会采取什么行动？
3. 列举建立一个虚电路的主要阶段。
4. X.25网络能否脱离PAD运行？
5. 如果进入帧中继接口的优先级数据流没有被平均强度限制，那么对于尽力服务的数据流会发生什么？
6. 假如一个用户在一个不支持通过电话网络自动调用的终端上工作，他如何才能使电话网络接入内建的PAD？
7. 假如你的公司需要把几个远地办公室网络接入到中心网络和另一个网络，你仅有安装同步19.2Kb/s的调制解调器的租用线路，那么你会选择以下那种技术：X.25、帧中继，或者ATM。请解释你所选方案的理由。
8. 令牌桶算法中哪一种功能，在漏桶算法中不支持？
9. 使用AMT网络传输语音时，使用哪一种服务类别比较有利？
10. 假如你需要传输三种有不同QoS级别的数据流类型，在两个ATM交换机之间每个方向你需要建立多少虚电路？
11. ATM专门用于数据流控制的是什么服务类别？为什么不使用其他类型的流量控制？
12. 假设你需要手工在两个企业级别的ATM网络间建立一个PVC，这两个网络是通过一个公用ATM网络连接的。你不需要依靠公用ATM网络管理员使用的VCI编号来制定你自己的VCI编号。你需要向公用ATM服务提供商要求那种类型的交换？
13. 假设你使用帧中继网的一个PVC来创建了远程网桥，用于连接两个局域网。两个网络之间的两台计算机间的NetBEUI会话经常断开。当计算机属于同一个网络时，没有这样的问题。这是什么原因？

练习题

1. 当两个终端节点交换两个TCP消息时（发送数据和收到应答），比较以下两种方式产生的帧的数量：两个节点通过一个X.25交换机相连，和通过帧中继交换机相连。
2. 在以下那种情况，使用帧中继网传递到目的节点的帧比例最高：当服务被基于CIR、B、Be参数请求，或者服务仅基于CIR，B_c（假设CIR和B_c在两个例子下相同）。假定帧中继网没有超负载，并且发送节点经常显著超过CIR值发送数据。
3. 假设帧中继交换机和IP交换机都基于同样的结构，处理器的时钟频率也相同。是否帧中继交换机能够提供更好的性能？说明你的理由。
4. 解决图21-12中的ATM网络数据流工程任务。你必须保证对于图21-13中提供的负载，所有的网络资源有最均衡的负荷。

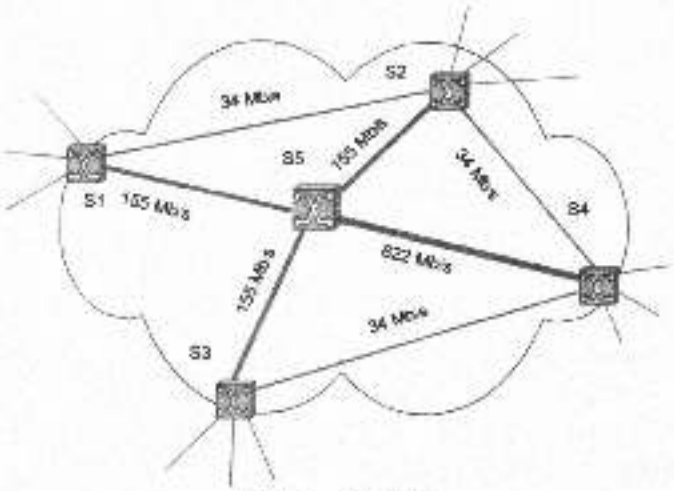


图21-12 ATM网络

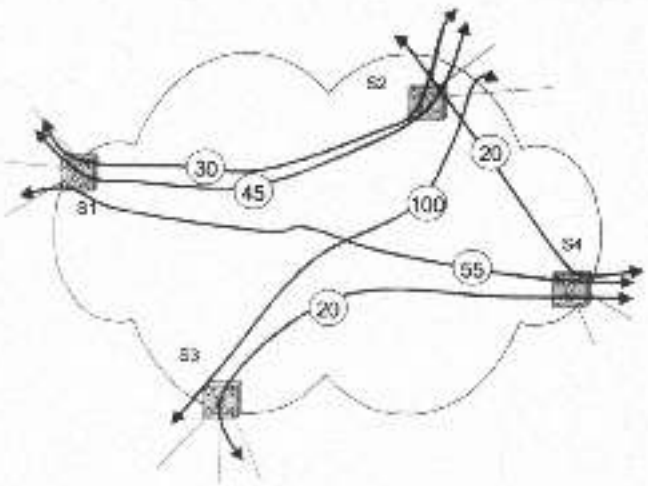


图21-13 提供的负载

第22章 IP WAN

22.1 引言

由于因特网的迅速普及，几乎每个WAN都有传送IP流量的能力。这意味着几乎所有的当代WAN都是IP因特网，而它们之间的所有差异都在IP层之下的技术。

IP WAN中的IP路由器直接由物理的点到点链路连接，这些IP WAN有着最简单的结构。在这种情况下，IP路由器连接退化的组成网络。根据对术语“网络”的常见解释，这些组成网络根本就不是网络，因为它们仅包括两个节点——相邻路由器的接口。在本书中，这种网络被称为“纯”IP网络，因为在由IP路由器形成的层的下面没有分组交换层。这意味着，IP路由器独自执行所有与分组转发相关的任务。

在20世纪90年代中期，IP WAN的多层结构变得最为流行。在这种WAN中，ATM和帧中继网络被用做分组网络。在两层上使用基于不同原理（数据报和虚电路）的分组交换网络使得IP WAN复杂而且昂贵。然而，传送多媒体信息的可能性、使用QoS和流量工程方法来平衡负载及优化网络资源使用，对这些缺点进行了补偿。

多协议标记交换（MPLS）技术是在IP与虚电路技术相结合的领域里的另一个创新。MPLS位于IP层与ATM、帧中继或以太网层之间，这样就将它们综合为统一的更有效的结构。

本章最后是对基于简单网络管理协议（SNMP）的网络管理系统的介绍：它不但被广泛用于控制IP路由器（这也是开发它的目的），还被用于控制任何类型的通信设备，从SDH或DWDM多路复用器到电话交换机。

22.2 纯IP WAN

22.2.1 IP WAN结构

为了提供一定范围的高质量服务，大多数大型WAN尤其是那些商业通信运营商拥有的WAN，是根据四层结构设计建造的（图22-1）。

下面两层与分组交换网络不相关。在传输网络的最低层，采用的是目前速度最快的DWDM技术。它创建了保证传输速度大于或等于10 Gb/s的光谱信道。SDH技术（和PDH访问网络）作用于下一层。使用这项技术，这些光谱信道的带宽被分为TDM的子信道，连接在分组交换网络的交换机（或者电话交换机）的接口。

基于传输网络，每个通信运营商都可以迅速组织起一个永久数字链路，这条永久数字链路位于覆盖网络的设备连接点之间，要么是电话网络要么是分组交换网络。

图22-1所示的WAN模型的最高层是由一个IP网络组成的。

说明 因为这些层执行的是OSI物理层的职能，所以不能把图22-1所示的各层与相应OSI模型的各层直接一一对应。

图22-1展示了如今用于建造传输网络的最具可延拓性的变体。这种构造包括DWDM层和SDH层。如今这种结构仅用于横跨国家或者整个大陆的最大型的地域网络。在很多小型的主干网中，没有DWDM层，甚至不使用SDH技术。通信运营商可能使用速度较慢、容错性较差但更经济的PDH技术来取代DWDM或SDH。

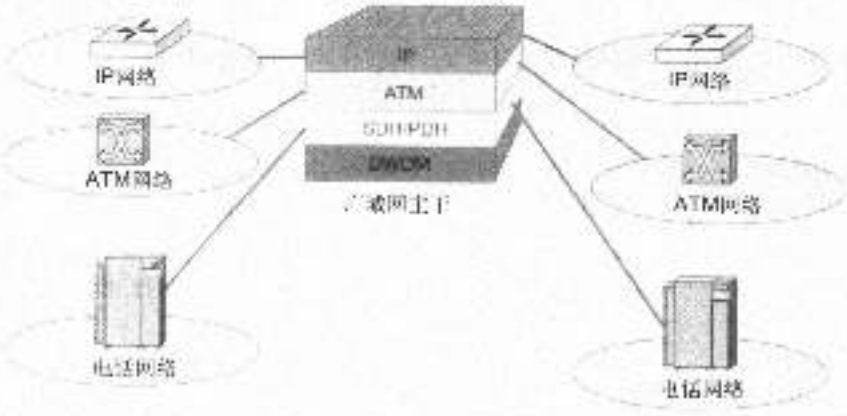


图22-1 当代WAN的四层设计

在更简单的情况下，根本就没有传输网络，在IP层之下可能就是ATM或者帧中继网络。这种网络的交换机使用电缆或者无线链路直接连接。虽然后一种方法要求的投资少一点，但是它的灵活性较差，因为它必须为每一个新设备建立一个新的物理链路。另一方面，一个带分支的传输网络的实用性允许在已有网络中通过构造多路复用器的交换矩阵或DWDM/SDH的交叉连接来创建新的信道。

在图22-1的设计中，第二层由ATM网络组成，它的主要目的是创建一个有*QoS*保证的永久虚电路（*PVC*）结构。ATM网络创建的这些PVC连接IP路由器的各个接口。对于IP流量的各种类别——平均速率、突发性、延迟级别和丢失级别，ATM网络都会形成一个单独的虚电路以保证服务该流量类别所需的QoS参数。使用ATM作为IP的基础技术不仅保证了用户流量所需的QoS，而且还解决了通信运营商在流量工程基础上的内在问题。该问题的解决保证了传输网络的所有物理链路的负载均衡。

从保证QoS参数中解放出来的IP层，执行它的传统职能——即，建造一个互联网并给终端用户提供IP服务。终端用户可以通过该WAN传送它们的IP流量或者使用这个网络连接到互联网上。

尽管这种多层结构较为复杂，但这种网络已经变得广为流行。对大的通信运营商和服务提供商来说，它们是实际上的标准。使用这个标准，他们可以提供复杂的服务，比如：IP服务、ATM服务，传统的电信服务还有租用数字链路。除了IP服务，这些服务的用户直接与需要的通信运营商网络的层次进行通信（例如：ATM、电话、SDH或DWDM）。

然而，在很长一段时间里，IP网络没有如此复杂的多层结构。传统的IP网络由直接通过通信链路连接的路由器组成，这些网络不提供QoS支持，因为20世纪80年代的应用产生的流量对延迟并不敏感。在多层IP WAN产生后，则必须区分这两种网络；因此，传统IP网络被称做“纯”IP网络。

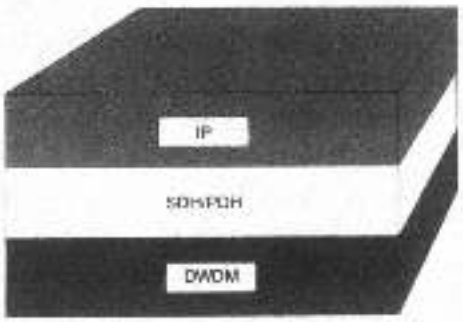


图22-2 纯IP网络的结构

纯IP网络也有它们的应用。它们的各个路由器在IP层之下没有分配交换网络（如ATM或帧中继），并且IP路由器之间利用点到点链路直接连接（物理连接或PDH/SDH/DWDM连接）。

图22-2展示了一个纯IP网络的结构。

在这些网络中，数据链路层以前一样由前两层的基础来建造。这些链路被IP路由器的接口直接使用而不用使用任何中间层。如果IP路由器使用在SDH/SONET

网络中创建的信道，则这个变体就叫做**同步光纤网上的分组**（packet over SONET，POS）。

在下列情况下，纯IP网络可以成功传送如今应用程序的延迟敏感流量：

- IP网络作用于负载不足模式下；因此，所有类型的服务都供大于求，不须在队列中等待。因此，这种网络不需要QoS支持。
- IP层使用集成服务（IntServ）或区分服务（DiffServ）机制保证QoS支持。

在纯IP WAN模式下，为了让IP路由器使用数据链路，在这些链路上必须运行某个数据链路层协议。

目前已经为点到点WAN连接开发了多个数据链路层协议。这些协议有内嵌的程序用于WAN中的操作，包括：

- 数据流管理
- 远程设备的相互识别。用于对付为了窃听而探测并重定向流量的“虚假”路由，常需要用这个特征来保护网络。
- 在数据链路层和物理层上的数据交换参数的协商。当两台设备位于不同的城市时，在开始为远程通信交互数据前，进行一些参数的协商，如数据域的最大容量（MTU），是非常有意义的。

如今，IP使用两个已有的点到点协议：高级数据链路控制（HDLC）协议和点到点协议（PPP）。同时还有一个被弃用的协议，串行线路网际协议（SLIP），它曾在很长一段时间里是个体用户经由电话线路访问IP网络所用的主要协议。现在，它已被成功运作于主干网和信道访问的PPP所取代。

除了这些为WAN点到点连接开发的协议外，WAN IP路由器在租用线路上还使用一种高速以太网变体。它们可能是快速以太网、千兆以太网或10G以太网。这些以太网协议不支持前面所列的那些程序。自然，这些程序对WAN很有用，然而，在LAN中以太网的流行暗示了它们受欢迎的程度。

22.2.2 HDLC族的协议

高级数据链路控制（The High-Level Data Link Control，HDLC）是为下列网络创建数据链路层的协议族：

- LAP-B——创建X.25网络数据链路层的协议
- LAP-D——用于ISDN网络
- LAP-M——用于同步和异步模式
- LAP-F——用于帧中继网络

关于HDLC应该注意的是它的复杂性。它可以运行在多个彼此不同的模式下，并且既支持点到点也支持点到多点连接。为了与工作站通信，HDLC还提供了多个不同的职能。对这个协议的操作细节感兴趣的读者可以在一些著名的流行书籍中找到相关内容。这里我们仅涉及这个协议族成员都支持的HDLC基本职能。

HDLC支持三种逻辑连接模式，它们的区别在于通信设备扮演的角色。我们仅讨论其中的一个——*异步平衡模式*（*ABM*），因为这个模式是IP路由器所使用的模式。在ABM模式中，两个设备有平等的权力。它们交换被分类为信息帧和响应帧的帧。

HDLC帧的格式见图22-3。

HDLC帧包含下列字段：

- 标志。HDLC帧封装在编码为01111110 的首尾标志之间。接收端使用这种标志来判定帧的起始点和结束点。这种方法省去了HDLC帧中帧长度字段的使用，但是这种方法的使用引起了另外一个问题——怎么区分值为01111110的数据字节和标志字节。**位填充**（bit-stuffing）技术在这里就有了用武之地。位填充技术只在帧的数据字段传送期间运行。如果发送端发现已经连续传送了5个1，它将自动插入一个0到传送的位序列中（即使紧跟在这5个1 的序列后面

的就是一个0)。这样，在帧的数据字段中就不会出现序列01111110。在接收端上也有一个类似的设计；然而它执行相反的功能。当在5个连续的1后发现一个0时，这个0将自动从该帧数据字段中删除。

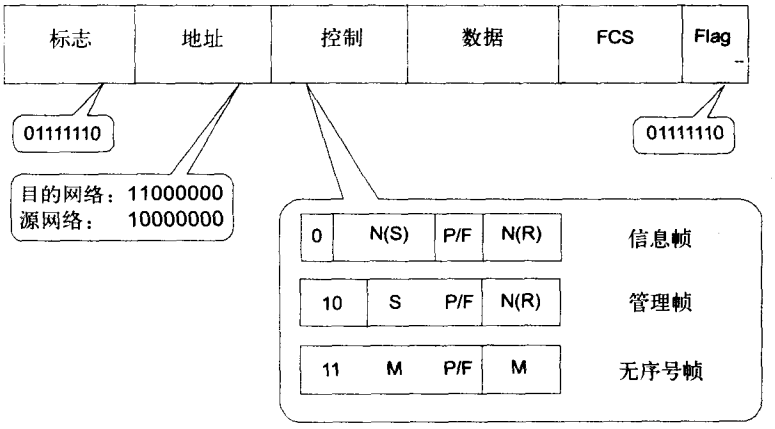


图22-3 HDLC帧

- *数据字段*。HDLC帧的数据字段是用来在网络上传输高层协议的分组。这些协议可能是IP、IPX、AppleTalk、DECnet；在少数情况下，如果它们把它们的报文直接封装到数据链路层的帧中时，也可能是应用层协议。在管理帧和一些无序号帧中，没有数据字段。
- *地址*。这些1或2个字节执行普通的从多个可能的设备中选择要求的那个设备的职能（只存在于点到多点的构造中）。在点到点的构造中，HDLC地址字段用来指出帧的传输方向——从网络到用户设备或者从用户设备到网络。显然，这个地址功能只有当经由UNI传送帧时才有意义。
- *控制字段*。这个字段长度为1或2字节，它的结构决定于要传送的帧的类型。帧的类型由控制域的起始位定义：0——*信息帧*，01——*管理帧*，11——*无序号帧*。在所有类型的帧的控制字段中都包含有P/F位。这个位（P/F代表选举/结束）在命令帧和响应帧里的用法不同。例如：如果接收端从发送端收到一个P位为1的命令帧，它必须把F位置1以作为响应帧立即回应。

现在再详细讨论一下不同类型帧的结构和目标。

- *无序号帧*既用于传送报错信息，还用于建立和终结逻辑连接。无序号帧用在为设备建立连接阶段。无序号帧的*M*字段定义了这些设备建立连接时所使用的命令类型。下面是一些命令示例：
  - *SABME*（*设定异步平衡扩充模式*）——这个命令是用于请求建立连接。扩充模式意味着使用2个字节的字段来控制其他类型的帧。
  - *UA*（*无序号确认*）——这个命令建立或者终止一个连接。
  - *REST*（*连接重置*）——这个命令请求终止连接。
- *管理帧*是用来在已建立的逻辑连接的环境下传递命令和响应的，包括发送损坏信息块的重传请求。管理帧包括以下几种：
  - *拒绝*（*REJ*）——这个命令被用于接收端拒绝应答。
  - *接收端尚未就绪*（*RNR*）——这个命令用于减缓发送给接收端的帧流。
  - *接收端就绪*（*RR*）——这个命令用于当从接收端到发送端之间没有数据流时的肯定应答。
- *信息帧*是用来传送用户信息。在信息块传送期间，根据滑动窗口算法给信息帧编号。

建立连接后，数据和肯定应答通过信息帧传递。逻辑HDLC信道是*双向的*，这也意味着信息帧和肯定应答可以双向传送。如果在相反方向没有数据流，或者如果有必要传递一个拒绝应答时，就要使用管理帧。

为了保证可靠性，HDLC使用了一个7帧（如果控制字段大小为2 字节）或127帧（如果数据字段为2字节）的滑动窗口，且用0到127的数字给它们循环编号。为了支持滑动窗口算法，从发送端发出的信息帧提供了两个字段：

- $N(S)$域指定发送帧序号。
- $N(R)$域指定期望收到的下一帧的序号。

为了说明它们的不同，假设站A发送了一个值为$N_A(S)$和$N_B(R)$的信息帧给站B。如果从站B发回的响应帧序号$N_B(S)$与站A所期望的序号——$N_A(R)$相等，则这次传输被认为是正确的；如果不等，那么站A丢弃该帧并发送序号为$N_A(R)$的REJ否认应答。收到该否认应答后，站B必须重传序号为$N_A(R)$的帧，并使用滑动窗口算法重传所有已发的序号比$N_A(R)$大的帧。

当接收端到发送器之间没有数据流时，就使用$N(R)$字段被置位的RR管理帧作为确定应答。如果滑动窗口机制不能控制帧流，则使用RNR命令要求发送器挂起传输直到收到RR命令。

22.2.3　点到点协议

点到点协议（Point-to-Point Protocol，PPP）是因特网标准。和HDLC协议类似，它也代表了整族的协议，实际上它主要包括：

- **链路控制协议**（Link control protocol，LCP）
- **网络控制协议**（Network control protocol，NCP）
- **多链路PPP**（Multi Link PPP，MLPP）
- **密码识别协议**（Password Authentication Protocol，PAP）
- **挑战握手识别协议**（Challenge handshake authentication protocol，CHAP）

说明　PPP被开发出来时，是以HDLC帧格式为原型并补充一些新的字段。PPP的字段被封装在HDLC帧的数据字段里。后来，出现了把PPP帧封装在诸如帧中继和其他WAN协议的帧里的标准。虽然PPP可以与HDLC帧一起运作，但是与HDLC相比，它不提供保证帧传输可靠性和帧流控制的程序。

PPP和其他数据链路层协议的主要区别在于它使用一个专门的**协商程序**（negotiation procedure）实现了不同设备的协同运作。在这个程序中，这些设备互换多种参数，如线路质量、识别协议和封装的网络层协议。

在一个企业网中，终端系统通常在临时存储分组的缓冲区大小、对分组大小的限制及支持的网络层协议方面都各有不同。连接终端节点的物理线路也随低速模拟线路到高速数据线路的不同而提供不同级别的QoS。

接受连接参数所依据的协议是链路控制协议（*LCP*）。为了处理所有可能的情况，PPP提供考虑了所有标准构造的默认参数集。当建立一个连接时，两台通信设备首先尝试使用这些默认设置。每个终端节点指定它的容量和要求。然后基于这些信息，采用双方都满意的连接参数。

协议之间的协商也许不能在某个具体参数上达成一致。比如，某个节点可能建议MTU为1 000字节。另一个节点可能会拒绝这个建议并建议为1 500字节。1 500字节可能又被第一个节点拒绝。这样，这个协商程序将在超时后终止。

PPP连接的一个最重要的参数是识别模式。对于识别，PPP提供的要么是默认的密码识别协议，要么就是询问握手识别协议，后者不经由通信链路传送密码，这样可以保证更强的网络安全性。用户可以添加其他的识别算法，还可以选择报头和数据压缩算法。

多协议支持是PPP支持多个网络层协议的能力，这也说明了为什么广泛使用的它是实际上的标准。与只能携带IP分组的SLIP或只能携带X.25分组的LAP-B相比，PPP既可以和数据链路层的LAN

协议一起运作，也可以与很多网络层协议合作——包括IP、Novell IPX、AppleTalk、DECnet、Xerox网络系统、Banyan VINES和OSI。

每个网络层协议都是使用适当的*NCP*单独配置的。这个配置程序首先确定这个协议会在当前PPP会话期间用到，然后对该协议的一些参数进行协商。参数中最大的值是为IP设置的。这些参数中包括通信节点的IP地址、DNS服务器的IP地址和IP压缩报头。对于每个配置高层协议参数的协议，除了通用的名字NCP外，还有其他的名字。这些协议是以对应的协议名称加上控制协议（CP）的缩写来命名的：IPCP、IPXCP等。

协议的可扩展性。可扩展性既可以解释为使用定制的用户协议取代默认的协议的可能性，也可以解释为把新协议添进PPP栈的可能性。这就允许了为每种情况创建一个最优的PPP配置。

PPP一个最吸引人的能力是它可以使用多个物理链路来创建一个逻辑信道，也就是众所周知的信道干线。这个能力是由一个被称为多链路PPP的附加协议实现的。

22.2.4 IP路由器使用的租用线

如图22-4所示的是IP路由器使用租用线路的方法。为了连接路由端口到租用线路，必须使用适当类型的DCE设备。这台设备被用来把一个路由器的物理接口转换为租用线路使用的物理层协议，例如，由V3.5到T1。

如果租用线路是个模拟线路，这时有必要使用一个调制解调器作为DCE设备。当处理数据线路时，则必须使用数据服务单元/信道服务单元设备（DSU/CSU）。

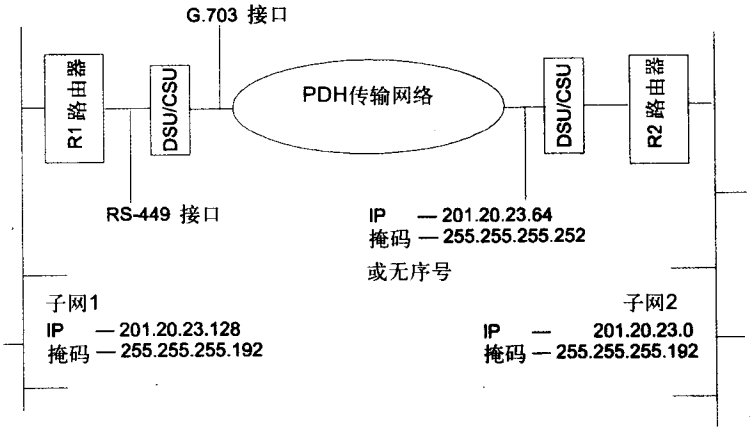


图22-4 使用租用线路连接IP网络

路由器的端口可能包含一个内嵌式的DCE设备。例如，当路由器使用一个SDH信道时，它通常有一个内嵌的端口，该端口带有一个特定STM-N速率的SDH接口。

内嵌的PDH/SDH端口有可能支持也有可能不支持这些技术的内部帧结构。当端口可以区分组成帧的子帧时——例如，E1帧的单独的时间槽或组成STM-1帧的单独的VC-12容器（2Mb/s）——可以把它们用作独立的物理子网，这时这些端口就被称为信道化端口。每个这样的信道被赋予一个单独的IP地址。如果不是这种情况，则整个端口被认为是有一个IP地址的物理链路。

在图22-4的例子中，两个路由器之间的连接是使用一个在PDH传输网络中创建的数字的E1信道建立的。为了连接信道，路由器使用了一个DSU/CSU设备，该设备带有一个内部的RS-449接口和一个外部的G.703接口，后者被定义为一个到PDH信道的访问接口。

在路由器连入租用线路和LAN后，必须对它们进行配置。租用线路是一个单独的IP子网，就像通过它连接的网络1和网络2两个LAN一样。这个子网也需要一个互联网管理员分配的IP地址。

在图22-4中，租用线路被分配了子网地址201.20.23.64。根据掩码255.255.255.252，这个子网有两个节点。

由租用线路连接的路由器的接口不一定非要被赋予租用线路的IP地址。这样的路由器接口叫做**无序号的**（unnumbered）。事实上，路由器会通过在租用线路上发送路由协议（RIP或OSPF）来获得它们。在租用线路上不使用ARP，因为在租用线路上硬件地址没有实际意义。

22.3　在ATM或帧中继上的IP

22.3.1　IP和ATM层间的通信

当在ATM或帧中继上建立IP网络时，ATM或帧中继网络工作于数据链路层和IP层之间。因为一个帧中继网络运行速率通常不超过2Mb/s[⊖]，而且延迟和延迟变化级别不在这个技术所支持的QoS变量清单内，所以中间主干网应用的最多的是ATM。

IP层和ATM层之间的通信如图22-5所示。

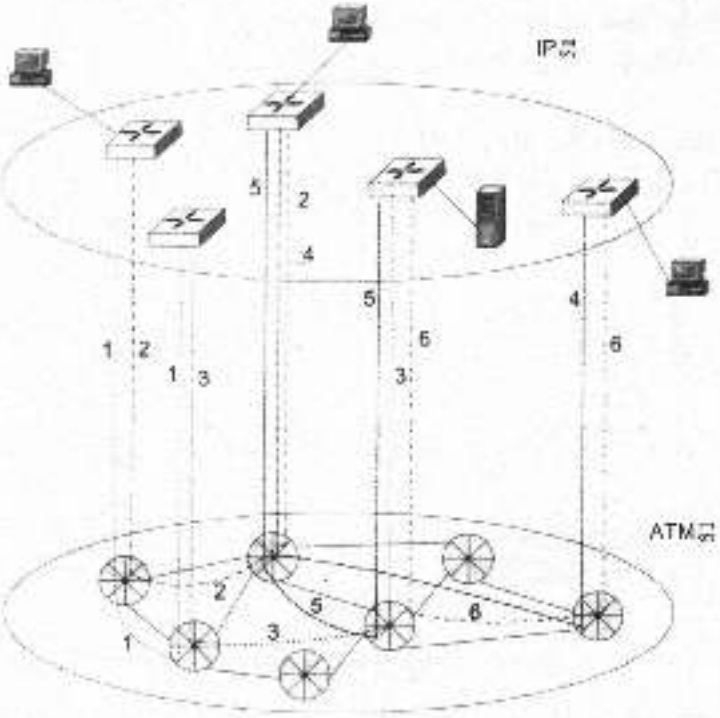


图22.5　IP和ATM间的通信

在一个ATM网络中，有六个连接到路由器端口的虚电路。每个路由器端口必须同终端节点一样支持ATM技术。虚电路建立后，路由器就可以像使用物理链路一样用它们来发送数据到下一个路由器（与虚电路有关）。

在ATM网络内部，虚电路用它自己的拓扑组成一个网络。图22-5所示的网络对应的虚电路拓扑见图22-6。这个ATM网络对IP路由器来说是透明的，因为它们不需要知道任何有关ATM交换机端口之间的物理连接的事。IP网络是有关ATM的一个覆盖网络。

⊖　最新版本的帧中继标准已把速率限制提高到622Mb/s。

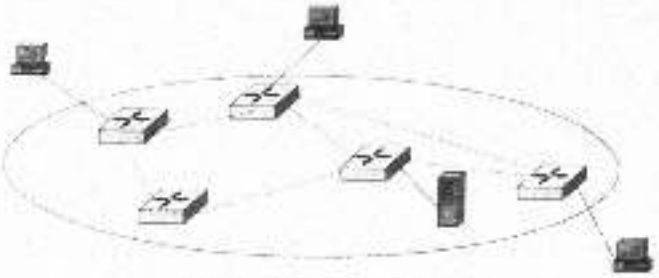


图22-6 路由器间的链路拓扑

22.3.2 配置路由器接口

为了让IP正确地工作，它必须知道邻居的IP地址和ATM虚电路号之间的映射，这些虚电路号可以用来到达要求的IP地址。换言之，它必须知道ARP表。在这种情况下，这种表不是通过发送ARP广播请求来自动创建的。相反，这种表必须手工创建。IP网络管理员必须为进出某接口的所有虚电路指定映射表，以此配置路由器的每个接口。同时，每个物理接口代表一个拥有独立IP地址的逻辑接口（或子接口）的集合。

例如，在Cisco系统路由器中，与VPI/VCI地址0/36的虚电路相对应的逻辑接口的配置如下：

```
pvc 0/36
protocol ip 10.2.1.1
```

实现这个指令后，路由器将知道什么时候需要发送分组到10.2.1.1的下一跳地址，它将不得不把分组分为ATM信元序列（使用ATM接口的SAR功能），然后都使用0/36 PVC发送它们。

如果一个多层ATM/IP网络必须传送不同类别的流量并观察每个类的QoS参数，那么相邻路由器一定是由多个虚电路连接起来的，每个类一个虚电路。必须为每个路由器指定一个指示每个传送的分组属于哪个级别的分组分类策略，每个类的分组都被送往适当的虚电路，这个虚电路保证该流量所要求的QoS参数。为了保证这些参数可以被观测到，有必要为ATM网络先执行流量工程，为了解决这个问题，需要为每个虚电路选择路由，这样它们就可以观察经由虚电路传送的流的平均速率。每个ATM交换机的每个接口的负载都不能超过预定义的阈值，以给每个流量类别保证一个可以忍受的延迟级别。

一个覆盖IP网络也可以使用交换虚电路模式来传送IP流量，这种模式适合于存在时间短的不稳定的流。为这些流创建PVC结构是低效的，因为它们大多数时间是空闲的。为了使路由器能够使用交换虚电路模式，就需要指定IP地址到路由接口的ATM地址的映射（例如，网络终端节点的ATM地址）。

与前面的情况类似，这个地址解决功能是由管理员手工执行的。Cisco路由器这种映射的一个可能的变体如下所示：

```
map-list a
ip 10.1.0.3 atm-nsap
33.3333.33.333333.3333.3333.3333.3333.3333.3333.3333.33
```

如果指定了这样一个映射，路由器为了发送一个分组到下一跳地址10.1.0.3，必须事先用Q.2931协议建立一个地址为33.3333.33.333333.3333.3333.3333.3333.3333.3333.3333.33的交换虚电路。然后，在自动从协议那收到一个VPI/VCI地址后，路由器就把源分组分裂的信元发送到该地址。收到这些信元下一个路由器的接口，把它们再组合成源分组并把这个分组传递给IP。

如果交换虚电路必须传送有一定QoS参数要求的流量，则这些参数就要依靠Q.2931协议传递，在为这个虚电路选择路由时，这个协议就会把这些参数考虑进去。

IP-over-ATM结构在通信运营商之间很流行，它们根据服务水平协定（SLA）提供它们的服务。

22.4 多协议标记交换

大多数网络专家认为**多协议标记交换**（MultiProtocol Label Switching，MPLS）是最有前途的传输技术之一。

MPLS下的多协议支持特点在于为它不仅可以使用TCP/IP路由协议，而且还可以使用任何其他栈的路由协议，如IPX/SPX。在这种情况下，MPLS可能会使用RIP IPX或NLSP来取代RIP IP、OSPF和IS-IS这些路由协议；LSR一般的体系结构还保持不变。

22.4.1 在同一设备中组合交换和路由

在同一设备中组合交换和路由这样的想法由*Ipsilon*在20世纪90年代中期第一次实现，开始制造组合的IP/ATM设备。这些设备实现了新的**IP交换**（IP switching）技术，它解决了使用前面所提到的交换路由技术所遇到的短期数据流传输低效的问题。为了在没有事先建立虚电路这个耗时的程序的情况下，传送短期流的分组通过ATM交换网络，Ipsilon建议在所有的ATM交换机中建造IP路由单元。这些单元使用RIP、OSPF或IS-IS这些标准的TCP/IP路由协议来构建路由表。

在Ipsilon网络中，短期流的传输按如下方式执行。从发送端节点来的分组抵达组合的IP/ATM设备，在那里它被拆分为ATM信元。此后，每个信元根据IP交换技术在IP/ATM设备间传递，从一台设备到另一台设备，然后沿着存于这些设备里的普通的IP路由表定义的路由一直传递到接收端。

典型的ATM技术的标准虚连接不是这样建立的；因此，短期IP流传送速度明显快很多。IP/ATM设备使用ATM传统的虚电路技术传送长期流。因为对IP和ATM协议来说，网络拓扑是一样的，所以对组合设备的两部分可以使用相同的路由协议。

为了实现这个技术，Ipsilon把专用的协议嵌入到IP/ATM设备中。这些协议负责识别数据流的持续时间，然后为长期流建立虚电路。这些协议曾以因特网草案的形式发布，但是还没有获得因特网标准的地位。

IP交换技术是为在通信运营商网络内部使用而开发的，它在与其他网络的边界处接收IP流量，然后以加速速率在主干网上传送。这是一个重要的特征，因为它允许这项技术独立于其他ISP使用，而且从外界角度看该运营商仍然是IP网络的一部分。

IP交换技术立即被那些通信运营商注意到，然后迅速流行起来。*Cisco System*通过创建更新的**标签交换**（tag switching）技术对Ipsilon的首创进行了进一步的发展。标签交换在合并IP和虚电路技术领域相当的先进。然而像IP交换一样，它也没有获得开放标准地位。

基于这些专用技术，包括来自不同公司的专家在内的IETF工作组开发了MPLS技术。

22.4.2 LSR和数据转发表

MPLS前身的主要原则被保留下来：MPLS使用路由协议去发现网络拓扑。它也使用虚电路技术在单个提供商的MPLS网络范围内转发数据。

不同技术的协议组合原则如图22-7和图22-8。第一个插图展示的是标准IP路由的简化的体系结构，第二个展示了支持MPLS技术的组合LSR设备的体系结构。

因为LSR执行了IP路由的所有职能，所以它包括了它的所有单元。为了支持MPLS职能，LSR也包括了与控制平台和数据平台职能相关的辅助单元。

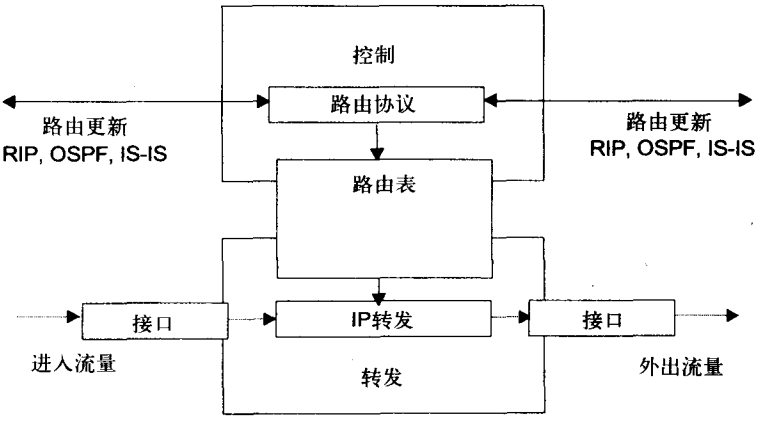


图22-7 IP路由器体系结构

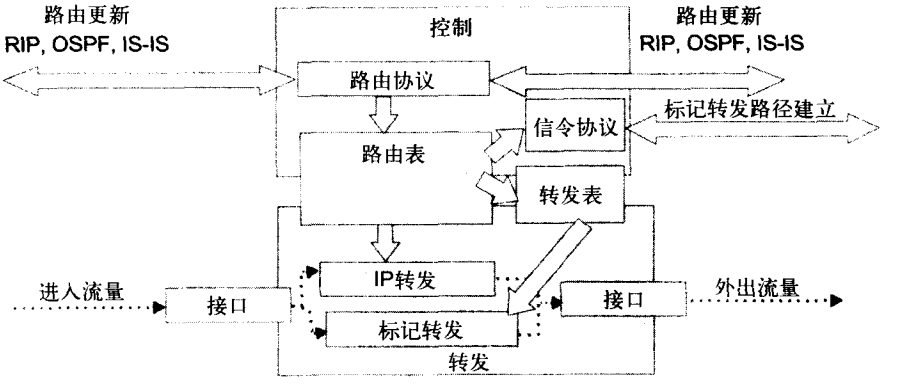


图22-8 LSR体系结构

标记转发单元（label forwarding unit）基于标记而不是目的IP地址传送IP分组。当做有关下一跳的决策时，标记转发单元使用**转发表**（forwarding table），它与基于虚电路的其他技术的转发表类似。表22-1为这种转发表的一个略图。

表22-1 MPLS转发表示例

输入端口	输入标记	下一跳	活动
S0	245	S1	256
S0	27	S2	45
…	…	…	…

你也许已经注意到这个表与第21章提到的一般转发表略有不同。这张表用*NextHop*字段取代了输出接口字段，同时还用*Action*字段取代了输出标记字段。在MPLS帧处理的大多数情况下，这些字段和一般转发表对应的字段的使用方法相同。这意味着*NextHop*字段包含了一个接口的序号，而帧就自动传递至这个指定序号的接口，包含在*Action*字段里的值则是新标记的值。这些字段的新名字意味着MPLS已经改进并推广了前述的所有的虚电路技术。所以，字段的值变得更通用。有时它们也用作其他用途，这将在本章后面提到。

每个LSR的MPLS转发表都是使用信令协议形成的，这个协议在MPLS里被称为**标记分发协议**（Label Distribution Protocol，LDP）。这个信令协议的功能类似于ATM和帧中继信令协议。

通过在LSR上建立转发表，LDP建立了虚拟路径，它在MPLS技术中有一个专门的名字：**标记**

交换路径（Label Switching Path，LSP）。

22.4.3　标记交换路径

图22-9展示了一个与多个IP网络进行通信的MPLS网络。

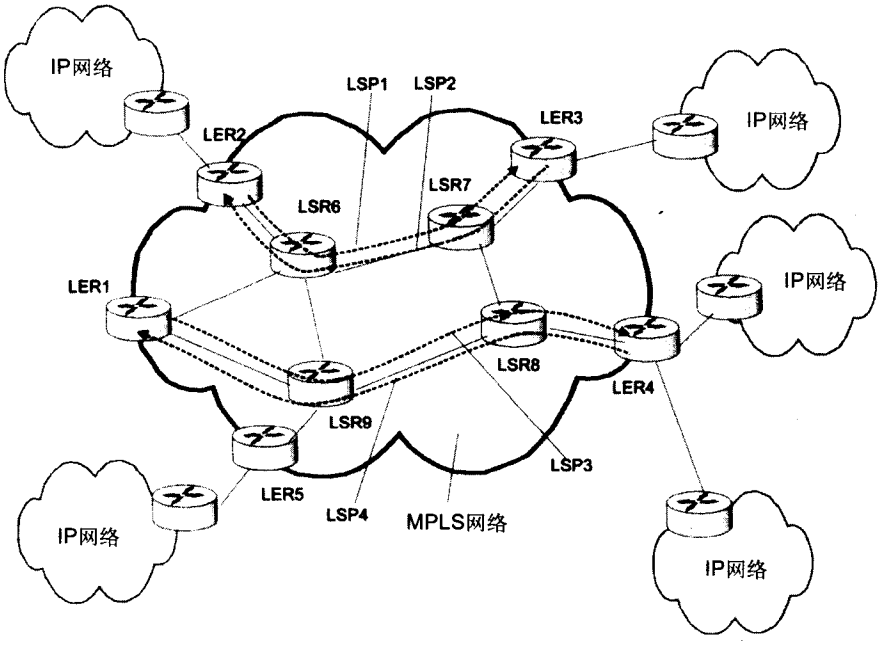


图22-9　MPLS网络

边界LSR在MPLS技术中也被叫做**标记交换边界路由器**（Label Switch Edge Routers，LER）。

LER从其他网络接收标准IP分组格式的流量，并给这个流量提供标记，然后沿着对应的LSP把它发送到另一个LER。每个中间的LSR基于流量的标记而不是目的IP地址转发分组。像使用虚电路技术的其他技术一样，标记有一个在每个LSR范围之内的本地值。当分组从输入接口传递到输出接口时，标记改变它的值。

在MPLS中，LSP是根据已有的互联网连接拓扑事先建立，与之形成对比的是IP交换技术是在创建长期数据流时建立LSP的。LSP是单向的虚电路；所以为了在两个LER间传送流量，就必须建立至少两个LSP，每个方向一个。图22-9展示了两对LSP，一对连接LER2和LER3，另一对连接LER1和LER4 。很明显，为了保证整个网络间的连通性，这是不够的。为了这个目地，需要在LER间有一个全连通的LSP拓扑。这个拓扑存在于现实世界的MPLS网络中，而在图22-9中没有显示它的原因仅是它太大了不便于用图形来表示。

LER输出（出口）从IP分组上删除了标记并把它以标准格式传递给下一网络。因而，MPLS技术的操作对其他IP网络来说仍然不可见的。

通常，MPLS网络实现上述分组处理算法的一个改进版本。这个改进版本的特点是标记的删除是由与最后相邻的（倒数第二）设备执行的，而不是在出口LER。事实上，在倒数第二台设备基于标记值决定下一跳时，已经不再需要MPLS帧中的标记了。最后一台设备如出口LER，将基于IP地址转发分组。对帧转发算法的这个小改动允许减少MPLS帧上的一个操作。没有它，出口LER必须首先删除标记，而且只能在那之后查询IP路由表。

22.4.4 MPLS头和数据链路技术

MPLS头由多个字段组成（图22-10）：

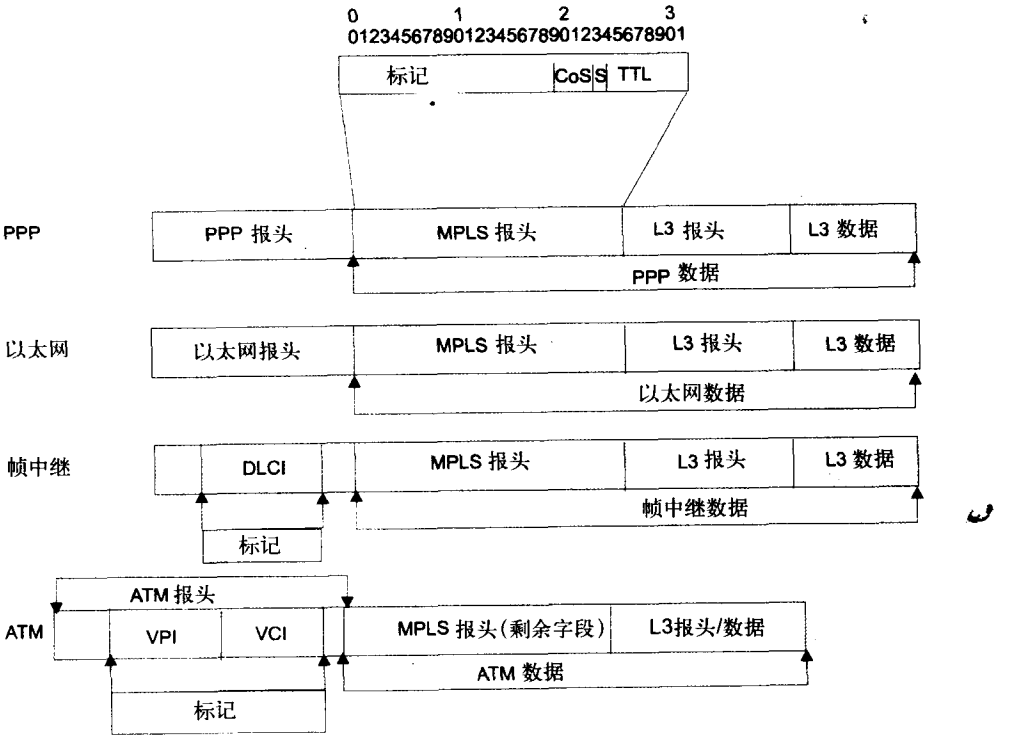


图22-10 MPLS标记格式

- 标记（*Lable*）——20位。标记字段用来选择对应的LSP。
- 生存时间（*Time To Live*，*TTL*）——8位，与IP分组的类似字段的完全一样。必需让LSR可以不需访问IP报头而仅基于MPLS报头里的信息就可以丢弃错误转发的分组。
- *CoS*（或实验性的）——3位。最初，这个字段是保留为将来使用的。最近，它主要用来为需要特殊QoS级别的流量指定类别。
- *S*——1位。这是标记栈栈底的指针。

标记栈的概念将在本章的下节详细讨论。现在，为了阐明MPLS和数据链路层技术之间的通信机制，我们来考虑MPLS报头仅包含一个标记的情况。

如图22-10所示，MPLS技术支持多种类型的帧：PPP、以太网、帧中继和ATM。这并不意味着这些技术里的任何一个都运行于MPLS层之下。而仅仅意味着MPLS使用这些技术中的帧格式来封装网络层分组（如今，这些分组实际上通常为IP分组）。

MPLS网络下的帧转发是在MPLS标记和LSP技术的基础上，而不是在寻址信息和帧格式为MPLS所使用的格式的那些技术所使用的技术的基础上执行的。例如，如果MPLS使用以太网帧格式，它就不使用目的和源MAC地址来转发帧，即使这些信息存在于以太网帧的对应字段中[⊖]。

⊖ 唯一的例外是在一个共享的媒介的变体下的MPLS/Ethernet，它的端口是依据“点到多点”设计连接的，同时输出端口的数值不能够提供足够的信息以决定下一跳。在这种情况下，为了分组转发，目的MAC地址是和标记一起使用的。然而，在如今的交换以太网统治的情况下极少遇到这样的一个变体，所以这里我们将不会提及这个变体。带输出接口值的MAC地址的使用解释了为什么输出接口字段被重新命名为下一跳（*NextHop*）。

对PPP、以太网和帧中继而言，MPLS报头被置于原有报头和第三层分组报头之间。ATM信元处理方式则不同。MPLS技术使用这些信元报头中已有*VPI/VCI*字段做为虚连接的标记。

*VPI/VCI*域仅用于存储标记字段。剩下的MPLS报头部分，包括*CoS*、*S*和*TTL*域，被存储在ATM信元的数据域中。当用支持MPLS的ATM交换机传送信元时不使用它。

稍后，当考虑这些不同时，将假设使用的是MPLS/PPP帧格式。

22.4.5　标记栈

标记栈（label stack）当前在是MPLS的特征之一。标记栈的概念是使用ATM所采用的*VPI/VCI*标记虚电路路径的概念的进一步发展。

标记栈允许创建一个任意层的聚合LSP系统。为了支持这个功能，沿着按层组织的路径传播的MPLS帧必须包含一个对应于路径层的MPLS报头的数值。每层的MPLS报头都必须有它自己的字段集合，包括*Label*、*CoS*、*TTL*和*S*。这个序列以栈的形式组织起来，所以通常都有一个位于栈顶的标记和一个位于栈底的标记。后者由属性*S* = 1表示。可以在标记上执行下列操作，它们在转发表中的Actions字段指定：

- *Push*——把标记压入栈。如果栈是空的，这个操作仅仅把一个标记赋给分组。如果栈已经至少有一个标记，那么新标记将将已有标记向下移，最后占据栈顶位置。
- *Swap*——用另一个标记取代当前标记。
- *Pop*——把栈顶标记弹出栈。这个操作的结果是，栈中其他标记都往上移了一级。

MPLS帧转发通常是基于栈顶标记执行的。考虑图22-11所示的一组LSP的MPLS域的转发。

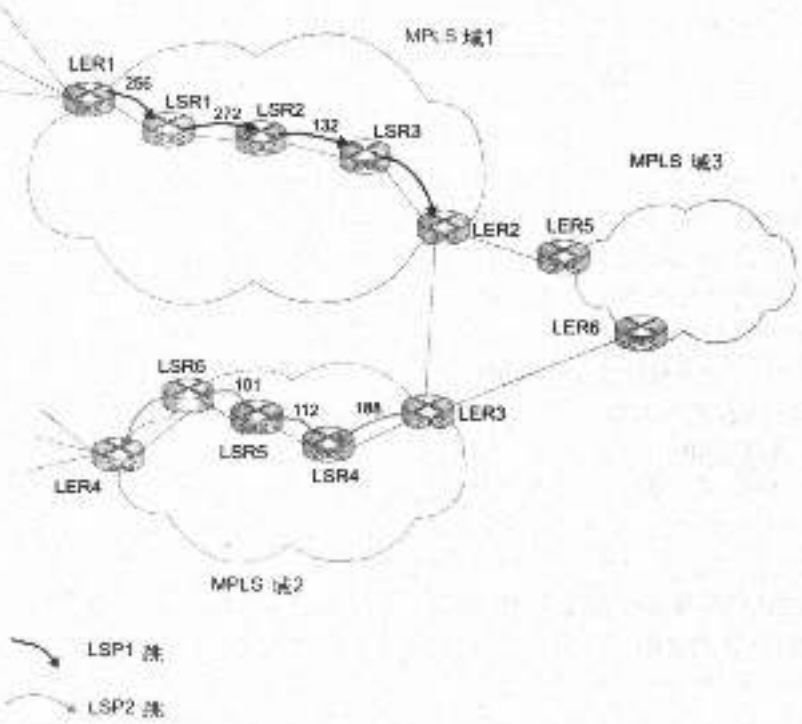


图22-11　MPLS域1和域2内建造的路径LSP1和LSP2

网络由三个MPLS域组成。图22-11展示了在MPLS域1内的路径LSP1和域2内的另一条路径LSP2。LSP1连接LER1和LER2，途经LSR1、LSR2和LSR3。假设256是LSP1的初始标记。这个标

记由入口设备LER1赋给分组。基于这个标记，分组被提供给LSR1。LSR1定义了标记的新值——272——基于它的转发表，然后把分组传递给LSR2的输入端。LSR2执行类似行为。它赋给分组一个新标记值——132——然后把它传递给LSR3的输入端。LSR3作为LSP1中的倒数第二个设备，执行Pop操作，把该标记从栈移出。LER2根据IP地址进一步转发该分组。

LSP2连接LER3和LER3，经过LSR4、LSR5和LSR6。这条路径由下面的标记序列定义：188，112，101。

为了不仅可以在每个域内而且可以在域和域之间（比如，在LER1和LER4之间）用MPLS技术传送IP分组，有两种解决方案可以选择。

- 第一个方案是在LER1和LER4之间建立一个连接LSP1和LSP2（在这种情况下形成一个单独路径）的一级LSP。当MPLS域属于不同提供商时，这个看似简单的解决方案是低效的。实际经验表明这种方案缺乏可延拓性，因为它不允许提供商单独作用。
- 第二种方案更有发展前景，它的特点在于使用一个多级的方法来连接两个MPLS域，这两个域可能属于不同的提供商，但这是不必要的。

在这个例子中，创建了一个第二层的LSP，LSP3来连接LER1和LER4。这个条路径定义了中间域跳的序列以取代每个域内的LSR之间的跳。因此，LSP3由跳序列LER1—LER2—LER3—LER4组成，并且没有定义域1和域2之间的精确路径。从这点来看，MPLS的多级方法在理念上接近于BGP，它也定义了AS之间的路径。

仔细考虑涉及两级LSP时，MPLS是如何运行的（如图22-12）。

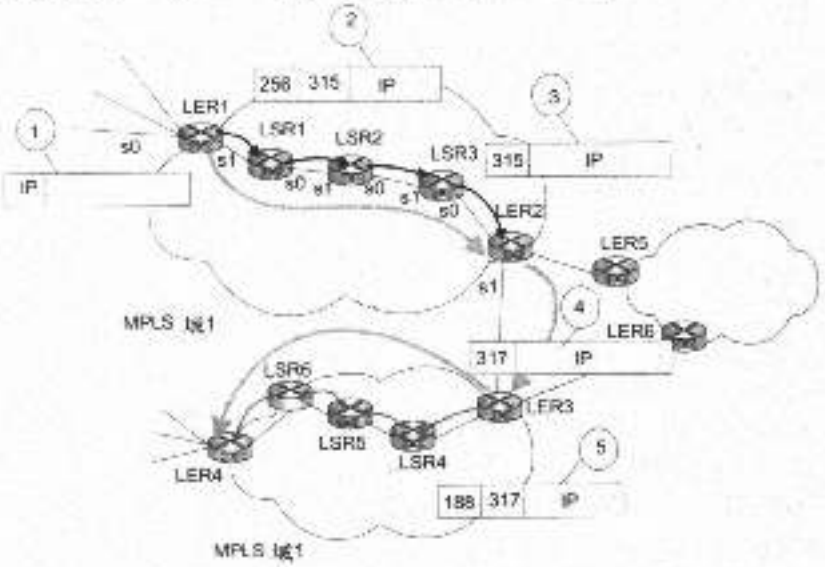


图22-12 在多级路径下使用标记栈

LER1是两条路径——LSP1和LSP3的起点。LER1转发表里的下列记录保证了这一点（表22-2）。

表22-2 LER1转发表的片段

输入接口	标 记	下一跳	活 动
...	...	...	...
S0	—	S1	315 Push 256
...	...	...	...

抵达LER1输入接口S0的IP分组被转发给输出接口S1。在输出接口，创建了一个由一个顶层标记315（LSP3）组成的MPLS报头。这个标记是当前居于栈顶的标记。然后，这个标记被压入栈底，同时另一个标记256（与LSP1相关）成为栈顶标记。最后，标记

值为256的MPLS帧到达LER1的输出接口S1，接着前往LSR1的输入端。

基于标记256，LSR1根据它的转发表处理这个帧，转发表包含下列记录（表22-3）。

位于栈顶的标记256被标记272取代。注意到LSR1 无视标记315的存在，因为当转发MPLS帧时，它不在栈顶上。LSR2 以同样的方式处理；它仅仅把标记值换为132，然后沿着路径把帧转发给下一个设备LSR3。

LSR3的操作与LSR1和LSR2有所不同，因为它是作为LSP1倒数第二个设备的LSR。它的转发表如下所示（表22-4）。

表22-3　LSR1 转发表的片段

输入接口	标 记	下一跳	活 动
…	…	…	…
S0	256	S1	272
…	…	…	…

表22-4　LSR3 转发表的片段

输入接口	标 记	下一跳	活 动
…	…	…	…
S0	256	S1	272
…	…	…	…

根据这条记录，LSR3对标记栈执行Pop操作。它删除了属于LSP1的标记132，所以LSP3的标记315成为了栈顶标记。对于路径上倒数第二个设备的LSR，这种行为是一种典型的设置，而不是多级路径组织。倒数第二个设备的弹出标记操作被称为倒数第二跳弹出（*PHP*）。

LER2 根据它的转发表的下列记录转发抵达它接口S0的帧（表22-5）。

LER2首先用317替换LSP3的标记值315，然后把标记317压入栈底，然后它把标记188置于栈顶。标记188是域2内的内部路径LSP2的标记。沿着LSP2的帧转发以类似方法执行。

这种两级模型可以很容易地扩展为任意级。

表22-5　LER2转发表的片段

输入接口	标 记	下一跳	活 动
…	…	…	…
S0	315	S1	317 Push 188
…	…	…	…

22.4.6　MPLS应用领域

在本章早些时候，你已经考虑过组成MPLS技术基础的主要原则。现在，MPLS技术已经实际应用于多个领域，在这些领域中，特殊的机制和为获得必需的功能性所需的协议实现了这些功能。MPLS的下列应用是最普遍的：

- MPLS IGP——这里，MPLS仅用于加速网络层分组转发。在这种情况下，分组沿着由标准内部网关协议（IGP）选择的路由传播，这也是在这个应领域使用MPLS名字的由来。
- MPLS TE——在这种情况下，根据修改的路由协议，MPLS LSP被选择来解决流量工程（*TE*）问题。MPLS TE不仅可以保证提供商网络的所有资源的负载的合理与均衡，而且为提供有QoS参数保证的传输服务打下了坚实的基础。
- MPLS VPN——MPLS这个领域的应用允许提供商在没有强制的数据加密条件下，根据流量隔离提供*VPN*服务。

在本章中，我们将考虑MPLS应用的前两个的领域的额外机制。MPLS VPN将在第24章和其他类型的VPN一起讨论。

注意，三种类型的MPLS可以在同一个网络内一起使用，给用户提供组合服务。

22.4.7　MPLS内部网关协议

MPLS IGP的主要目的是通过用交换取代路由来得到经由提供商网络的加速分组转发。所以MPLS的这个应用领域也被叫做**加速MPLS交换**（accelerated MPLS switching）。

当使用MPLS IGP时，LSP是根据已有的IP网络拓扑创建的，它们不依赖于这些网络之间的流量强度。这个特征见图22-13。

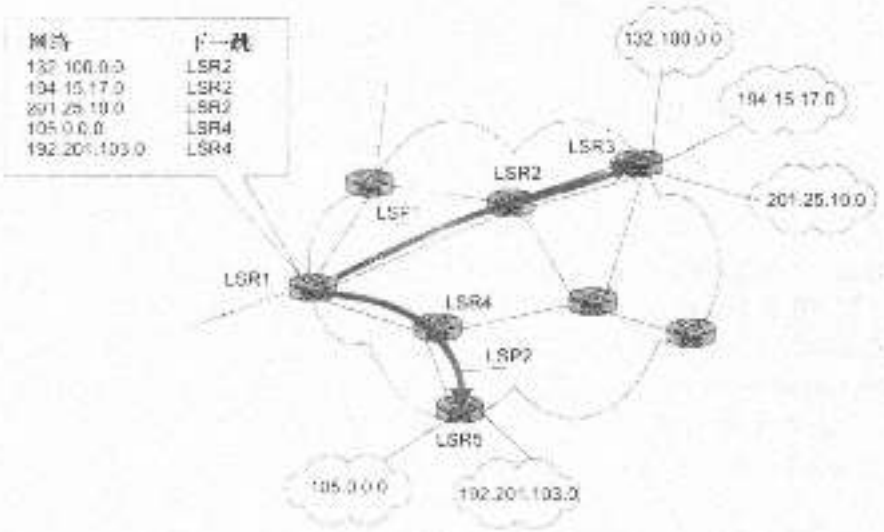


图22-13 使用LDP建立LSP

MPLS IGP的LSP是事先创建的，因此，当一个数据流产生时，传送该数据流的LSP必须已经存在。IGP LSP是自动建立的，网络管理员无需介入此程序。建立LSP所需的源信息是从MPLS提供商网络中的所有LSR的路由表处收集的。对于每一个在任意一个路由表中注册过的目的网络，都要创建一个条LSP。为了减少这些路径的数量，每个LSP都要可以服务前往多个目的网络的流量。这意味着IGP LSP服务聚合流量。同时聚合属性是与流量经过的提供商网络的沿途路由器相匹配的（完全或部分）。

网络的所有LSR都支持标记分发协议（label distribution protocol，LDP），它属于信令协议。每个LSR也必须支持一个标准的IGP，如RIP、IS-IS或OSPF。

作为路由协议作用的结果，或者网络管理员手动设定路由表后，一条关于目的网络的新记录就会出现在LSR路由表中。这时在提供商网络中还没有建立LSP，在这种情况下，这个LSR自动启动使用LDP建立新LSP的程序。

为了建立一条新LSP，发起的LSR使用了基于路由表的标准分组算法。例如，假设LSR1探测到在它的路由表中出现了一条新记录，其目的网络是132.100.0.0，指定的下一跳为LSR2。同时因为转发表没有对应的记录，所以没有到这个网络的虚拟路径。在这种情况下，LSR1发起虚拟路径创建工作。为了实现这个目的，它发送*LDP*请求报文给设备LSR2。在这条报文中，它指定需要建立的新LSP的目的网络的IP，LSR2收到这条报文后，根据自己的路由表和转发表的信息对它进行处理。如果LSR2也探测到它没有到网络132.100.0.0的LSP，它就传送LDP报文给下一个LSR，在LSR2的路由表中该LSR就被指定为到网络132.100.0.0的下一跳。在图22-13所示的例子中，LSR3扮演这台设备的角色，在它那LSP必须终止，因为下一跳已经超出提供商网络的范围。

说明 有个可能会问到的问题：LSR3是怎么知道它是该提供商网络的到网络132.100.0.0的路径上的最后一个LSR？注意到LDP是一个面向连接的协议，当建立一个逻辑连接时，LDP可以自动使用设备识别，所以LDP会话仅建立在同一个提供商的设备之间，该提供商给所有基于它的网络的LSR提供相互识别所需的信息。

探测到在到网络132.100.0.0的路径上的是一个LER后，设备LSR3给创建的LSP赋予一个它的输入接口S0没有使用的标记。然后，它发送一个*LDP*广告报文通知LSR2这个事件。然后轮到LSR2，它赋给这个LSP一个它的接口S0没有使用的标记，并且发送一个适当的*LDP*广告报文给设备LSR1。之后，从LSR1到网络132.100.0.0的一条新LSP就建立起来了，然后分组开始基于标记和转发表而不是IP地址或路由表在这条路径上传送。

为每个路由建立一个单独的到目的网络的LSP是不合理的。所以LSR试图建立聚合LSP，沿着它分组可以传送到多个目的网络的。例如，LSR1 可以不仅传送经由LSP1到网络132.100.0.0的分组，还传送到网络194.15.17和201.25.10.0的分组，因为到这些网络的路径都在该提供商的MPLS网络范围内。

为了传送发给网络105.0.0.0和192.201.103.0的节点的分组，LSR1有另外一条路径LSP2。使用LDP，它不但可能聚合从输入LER到输出LER的整个LER序列都一致的路径，而且还可能聚合仅有部分LSR相同的路径。所有有相同下一跳的目的网络的地址组成了当前LSR的所谓的**转发等价类**(forwarding equivalence class，FEC)。

通常，路由表所含的记录比转发表多，所以MPLS IGP只能通过减少路由表大小来加速分组转发。对于大的主干网，这个差值尤其值得注意，那里的路由器可能要处理包含几万条记录的路由表。

加速分组转发的另一个因素是缺少给路由器替换数据链路层帧的场所，这是IP技术所特有的。

22.4.8 MPLS流量工程

MPLS TE根据第7章所述的流量工程原则，执行创建有带宽保证的LSP的功能。

这是它与MPLS IGP的主要差异，MPLS IGP是在已知网络拓扑的基础上建立LSP的并且它忽略流量。

与MPLS IGP不同的另一个显著特征是，MPLS TE不是自动建立LSP的。TE LSP也被称为**TE隧道**（TE runnel），是由网络管理员主动建立的。就这方面而言，TE 隧道类似于ATM和帧中继这类技术的PVC。

MPLS TE支持两种类型的隧道：

- **严格的**（Strict）——定义两个LER之间所有的传送节点。
- **松散的**（Loose）——仅指定LER之间的部分传送节点。所有其他传送节点由LSR选择。

图22-14展示了以上两种隧道。

在这个例子中，隧道1是一个严格的隧道。当创建这样的隧道时，网络管理员不但要指定隧道的起始和末端节点，还要指定每个传送节点（如，整个地址序列：LER1、LSR1、LSR2、LSR3、LSR4和LER3）。因此，管理员通过选择通常有足够的可用带宽的路径来手动解决流量问题。当创建隧道1时，管理员不但要指定需要的地址序列，而且还要指定需要的带宽。虽然路径是在离线模式下选择的，但实际上隧道1上的所有设备都要检查是否能够提供请求的带宽，只有所有的设备都给出肯定的答复时隧道才建立好。

隧道2是松散的。当建立这个隧道时，管理员仅指定隧道的起始和末端节点（如，只有节点LER5和LER2）。隧道2的起始节点自动寻找传送节点LSR4和LSR2（如，由LER5寻找）。然后LER5使用信令协议通知这些节点和末端节点建立隧道的必要性。

独立于隧道类型，隧道通常有一个用于保留带宽这样的参数。在图22-14中，隧道1为流量保留了10Mb/s的带宽，隧道2保留了36Mb/s。这些值由管理员定义，MPLS TE技术对这些值没有影响。它仅执行请求的保留。大多数情况下，管理员在网络流量度量的基础上对隧道保留带宽做出评估，倾向于它的变化和自己的直觉。MPLS TE的有些实现允许在对流经隧道的实际流量的自动

流量的基础上对保留带宽做自动校正。

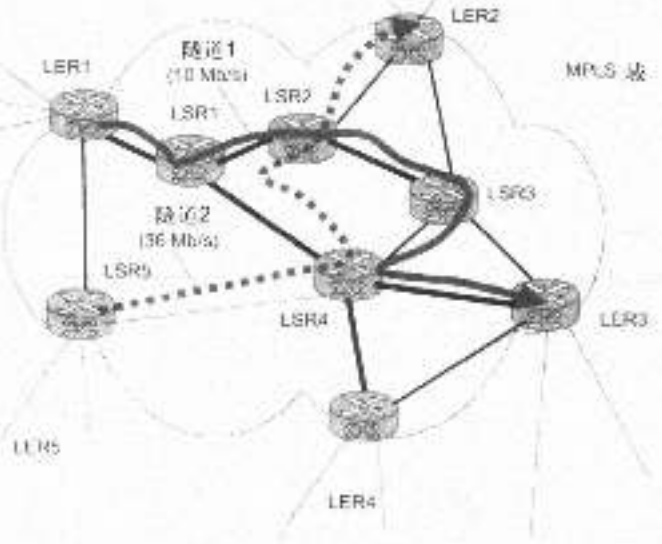


图22-14 两类MPLS TE隧道

然而，在具有MPLS的能的网络中建立TE隧道，并不意味着流量就是经由这个隧道传送的。这仅意味着有以不超过保留的值的平均速率传送流量通过这条隧道的可能性。为了在实际上通过这条隧道传送数据，管理员必须手动执行另一条程序：为隧道起始LER指定特殊条件，规定哪些分组必须通过隧道传输。有许多的这样的条件。所有的传统属性都可以用作聚合流的属性，包括IP目的地址、IP源地址、协议类型、TCP/UDP端口号、引入流量的接口编号和DSCP或IP优先级值。

因此，为了保证流平均速率不超过保留值，LER必须先执行**流量分类**（traffic classification），然后是**监控**（policing）。这时，穿过源自TE隧道的初始标记来**标识**（mark）分组以使用MPLS技术传送流量穿过网络。在这种情况下，在分派这些路径时执行的计算将提供想要的结果，即，网络资源平衡和对每个流有保证的平均速率的观测。

我们还没有考虑网络LER和LSR为了选择检查的隧道，检查组织严格的隧道的可能性或建立一个隧道所使用的具体的协议集合。

MPLS TE技术使用特殊的路由协议扩展来选择和检查路径，这些协议运行于链路状态算法的基础之上。目前，这些扩展已经标准化为OSPF和IS-IS协议。

为了解决流量工程问题，OSPF和IS-IS协议包含了多种新的广告类型，用它们在网络发布关于每个链路的额定的和无限制的带宽（可供流量工程流使用）的所有信息。因此，在每个LER或LSR拓扑数据库中创建的结果网络图的边，都将标以这两个额外的参数。在除了决定流量工程路径所需的流的参数外，还拥有这样一个可以支配的图后，LER就可以找到一个合理的方法来满足第7章所阐述的网络资源利用的要求。大多数情况下，这种决策是在最简单标准的基础上做出的，这个标准是在最简化的路由选择之上的最大资源利用，这意味着对所有可能路径，该路径最优化标准可表示为$\min\{K_{max}\}$。

一般而言，管理员必须为不同的聚合流建立多个隧道。为了简化最优化任务，这些隧道的路径选择是一个一个地顺序执行的，并且管理员仅依靠他的直觉来决定这个过程中每一个步骤的顺序。显然，流量工程路径的顺序选择降低了QoS，因为如果同时考虑所有的流，则有可能发现一种

更合理的资源分配。

示例　在图22-15所示的MPLS TE2的例子中，限制为允许的网络资源利用率的最大值为0.65。在变体1中，方案1是考虑到下列流的顺序建立的：1→2→3。对第一个流来说，选择路径A-B-C是因为它满足限制条件（路径上所有的资源，包括链路A-B和A-C的使用，对应的路由接口利用率为50 / 155 = 0.32）。另一方面，该条路径也是最短路径（65 + 65 = 130）。第二个流也选择了同一路径A-B-C。在这种情况下，限制仍然满足，因为利用率为(50 + 40) / 155 = 0.58。第三个流沿路由A-D-E-C发送，它使用了链路资源A-D，D-E和E-C（利用率为0.3），变体1可以认为是满足的，因为任何网络资源的使用都没有超过0.58。

然而，如变体2所示，还有一个更好的方案。在这个例子中，流2和流3沿路径A-B-C方向，流1沿路径A-D-E-C方向，路径A-B-C上的资源利用了0.45，下面那条路径利用了0.5，这意味着更均衡的资源利用。同时，整个网络的最大资源利用率没有超过0.5。当同时考虑三个流的$\min\{K_{max}\}$限制及以2-3-1的顺序依次考虑流时就可以获得这个变体。

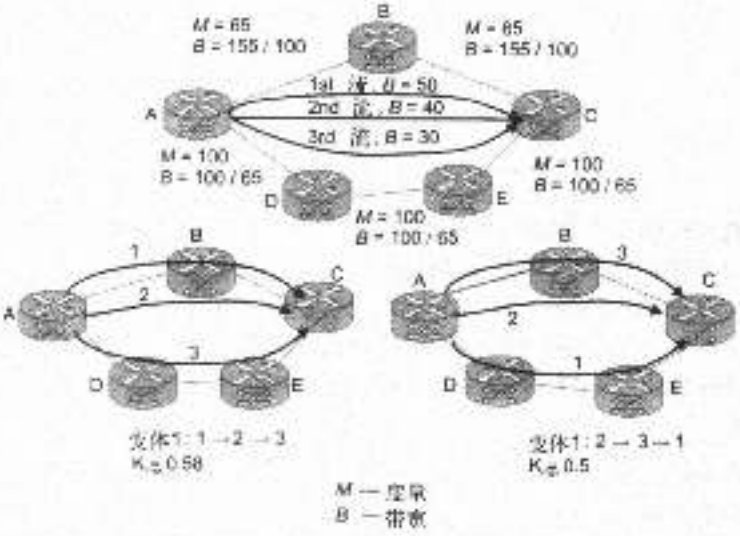


图22-15　流量工程在流选择顺序上解决方案的质量依据

然而，如今的MPLS TE设备使用的是顺序考虑流技术的变体，这种变体更易于实现，也更接近OSPF和IS-IS中使用的寻找到目的网络的最短路径的标准程序（没有限制时，最短路径集合的解决方案不依赖于考察网络的考虑顺序）。此外，当情况改变时（出现新的流或者已有流的速率发生变化），它有可能找到仅供某一条流使用的路径。

原则上，通过网络外部的自治的系统能为流的集合找到一个最佳方案。这可能是一个复杂系统，它包含一个不仅能考虑流的平均速率，而且还能考虑它们的突发性的仿真模型子系统。这个仿真器不仅可以评估资源利用，而且还可以评估QoS结果参数，如延迟和丢失。在找到最佳方案后，可以在一条一条地寻找路径的同时对它进行修改。

在MPLS TE技术中，充分利用了关于发现合理路径的信息。这意味着系统不仅如IP路由那样"记得"第一个传送节点，而且还"记得"路径上的所有节点，包括起始节点和末端节点。这种路由被称为源路由。因此，足以把寻找路径这个任务委派给网络LER，LSR可以仅提供它们有关链路带宽预留的当前状态的信息。

路径找到后，就必须建立起来，无论是由LER发现的还是在离线模式下发现的。用于这个目的，MPLS TE使用了资源预留协议（RSVP）的扩展，在这种情况下它称为RSVP TE。RSVP TE

报文根据路由的IP地址信息在LSR之间传送。当建立新路径时，信令报文指定保留的带宽和路径上的地址序列。每个LSR收到这个报文后，从对应接口可用带宽池中减去要求的带宽，然后在路由协议各自的广告中声明这个余数，如CSPF。

在结束这节之前，考虑一下MPLS TE和QoS技术之间的相互关系。如上所述，MPLS TE的主要目的是使用MPLS功能来实现服务提供商的内部目标，即，所有网络资源的负载平衡。此外，它还给运输服务提供有保证的QoS参数打下了基础，因为如果没有超过网络的最大利用率，就可以经由TE隧道进行流量传输。正如第7章所说的，资源利用对排队进程有决定性意义。因此，经由TE隧道传送的流量保证了某个已承诺的QoS级别。

为了保证每个流量类别的多种QoS参数，提供商必须为该流量类别的隧道创建一个独立系统。同时，必须执行对延迟敏感流量的预留，以便网络最大利用率在0.2～0.3范围之间。否则，分组延迟和它们的变化将超过允许的界限。

22.5 网络管理

22.5.1 网络管理系统的目的

任何复杂的远程电信网络除了标准网络操作系统提供的管理工具外，还要求有专门的管理工具。这是因为对于网络运行来说，大量各种各样的通信设备的正确运行是至关重要的。如果没有一个可以自动收集每个集中器、交换机、路由器或多路复用器的状态信息并把这些信息传递给网络操作员的中心系统，那么大规模网络的分布式天性将决定了它的运行是不可能的。第一个被广为使用的网络管理系统是SunNet Manger，由SunSoft于1989年发布。SunNet Manger系统主要是面向通信设备和网络流量控制的。这些是大多数网络专业人士提及**网络管理系统**（network management system）时想到的功能。

通常，网络管理系统运行于自动模式，执行与自动网络管理相关的最简单的任务。管理员则在网络管理系统收集和准备的信息的基础上做出最复杂的决策。

网络管理系统通常都是硬件与软件的复杂组合。所以，肯定有一定的限制，在该限制范围内它们的应用为合理的。在一个小网络中，有可能为像VLAN交换机这样的最先进的设备运行独立的小程序。通常，制造商为任何需要一个复杂配置程序的设备提供了一个自治的配置、维护和管理程序。然而，随着网络的成长，可能会出现一个关于把所有独立的配置和管理程序组合为统一的管理系统的问题。为了解决这个问题，你可能不得不丢弃这些独立的工具，而用一个**综合网络管理系统**（integrated network management system）替代它们。

22.5.2 网络管理问题的功能组

无论被管理的对象是什么，我们都希望有一个这样的管理系统，它能够实现国际标准定义的一些功能，能归纳在不同领域使用管理系统的经验。ITU-TX.700建议和相似的ISO-7489-4把网络管理系统的任务分为下列5个功能组：

- 配置管理任务既包括配置整个网络还包括配置网络元素。对路由器和多路复用器这样的网络元素而言，这些任务包括定义网络地址、标识符（名字）和地理位置。对整个网络来说，配置管理通常由构建展示网络元素间实际链路的网络图开始，如建立新的物理的或逻辑链路和改变交换表或路由表。
- 故障管理包括检测、定位和删除网络故障和它们的影响。
- 性能管理任务与分析积累的统计学信息有关，并使用它作为评估下列参数的基础：系统响应时间、连接两个网络用户的实际或虚拟信道的带宽、在独立的网段或链路中的流量速率、在网络上传送信息时数据损坏的概率及整个网络的可用性或它的特殊传输服务的可用性。性能的结果

和可靠性分析允许网络用户和网络管理员（或者服务提供商）之间缔结的SLA可以被观测和监控。没有对性能和可靠性的分析，提供公共网络服务的服务提供商或某公司的信息技术部门就可能不能控制或确保提供给终端用户需要的QoS级别。

- 安全管理假设对网络资源（设备和数据）的访问是被控制的，并且在数据存储期间或在网络上传输期间数据的完整性是有保证的。安全控制的基本元素是识别用户、赋予并检查网络资源的访问权利、分发并维持密钥、管理特权等。这个组的功能通常被实现为专门的软件产品（如Kerberos识别和授权系统、防火墙和数据加密系统）或被嵌入操作系统和系统应用程序中。
- 计费管理包括对多种网络资源的使用计费，如链路、信道、设备和运输服务。因为不同提供商使用不同计费系统和不同的形式的SLA，所以商业系统和网络管理平台，如HP OpenView，通常不包含这一组的功能。取而代之的是，它通常实现在为个体客户开发的自定义系统中。

说明　虽然OSI模型在控制的对象之间没有不同，这些对象包括：链路、段、交换机、路由器、调制解调器、多路复用器、计算机硬件和软件，但实际上把管理系统按所控制对象类型划分的情况是很普遍的。这些经典的网络管理系统如SunNet Manager或HP OpenView，都只控制公司网络的通信对象，如交换机和路由器。当需要管理计算机及它们的硬件和软件时，管理系统通常被称为**系统管理系统**(systems management system)。这种系统通常自动收集网络上计算机的信息，然后创建关于具体的硬件和软件的记录，并把它们存入到专门的数据库中。一个系统管理系统可以在中心安装和管理运行于文件服务的应用，还可以远程控制计算机、操作系统和DBMS的最重要的参数（处理器或RAM使用，页面故障强度等）。一个系统管理系统允许网络管理员以仿真模式远程控制计算机。这些系统的例子有：微软系统管理服务器、CA Unicenter和HP操作中心。实践表明，再过一些年，网络管理系统趋向于把系统管理系统综合为统一的网络管理产品，如CA Unicenter。

22.5.3　网络管理系统的体系结构

任何网络管理系统的基本元素都是管理者—代理—被管理对象方法（图22-16）。基于这个方法，可以构建包含任意数目的管理者、代理和多种资源的任意复杂性的系统。

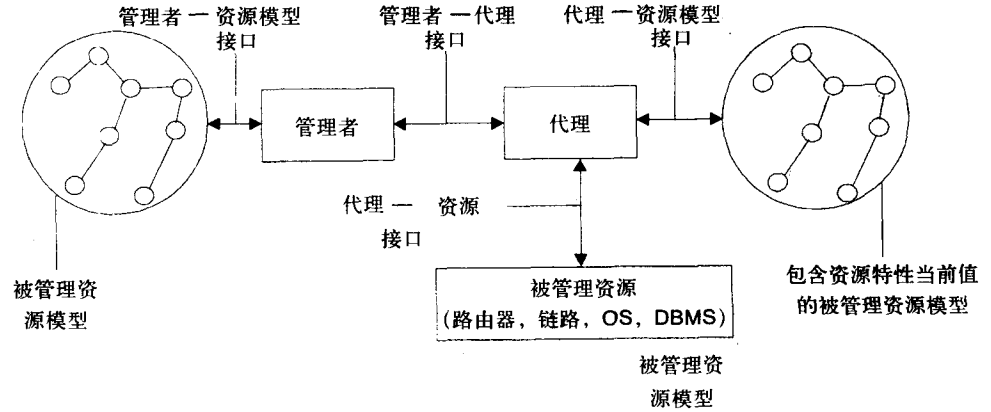


图22-16　代理，管理者和被控对象间的相互作用

为了自动控制网络对象，创建了一个被管对象的特殊模型，称为管理信息库（*MIB*）。MIB仅反映控制对象所需的对象特性。例如，一个路由器模型通常包括这些特性：端口的个数，它们的类

型，路由表，经由这些端口传递的数据链路层、网络层和运输层的帧和分组的数量。管理者和代理操作于被控对象的同一模型。然而，它们使用这个模型的方法有重大区别。

代理（agent）把被控对象特征的当前值填入它的MIB，然后**管理者**（manager）从MIB检索数据。根据这个信息，它得知可以向代理请求哪些特征以及可以控制对象的哪些参数。因此，代理被用作被控对象和管理者之间的接口。它只向管理者提供MIB提供的数据。

管理者和代理使用标准应用层协议在网络上通信。这个协议允许管理者请求存于MIB的参数，并给代理提供它控制这个对象所需的信息。通常，管理者运行于一台单独的计算机并与多个代理相互作用。

代理可以嵌入到被控设备中或运行于连接到被控设备的单独的计算机上。为了获得需要的关于某个对象的信息，并发送指令给那个对象，代理必须能与之通信。然而被管理对象的多样性不允许代理和对象之间的合作方法标准化。开发者在把代理嵌入通信设备或操作系统时解决这个问题。代理可能会配备用于收集信息的特殊的传感器——比如，中继联系传感器或温度传感器。代理也可以有不同级别的智能。例如，它们可以为了给通过某个设备的分组和帧计数，提供所需的最低级的智能。有些代理也可以提供足以在紧急情况时执行控制指令序列，建立时间依赖，过滤错误信息等的高级智能。

控制被分为两类：**带内**（in-band），它的控制报文和用户数据经由同一信道传送；**带外**（out-of-band）传递它的控制报文的信道独立于传送用户信息的信道。例如，如果有关管理者和嵌入路由器中的代理之间的通信协议的报文和用户数据经由同一网络传递，那么这是带内管理。另一方面，如果管理者控制一个运作于FDM技术基础之上的传输网络交换机通过一个独立的X.25网络，同时代理也连接到该网络，那么它有带外控制。带内控制更经济，因为它不需要为了传送控制数据而创建单独的基础设施。另一方面，带外方法更加可靠，因为当主通信链路的某些网络元素失败和设备不可用时，它允许网络设备被管理。

管理者-代理-被管理对象方法考虑了有复杂结构的分布式管理系统。(图22-17)。

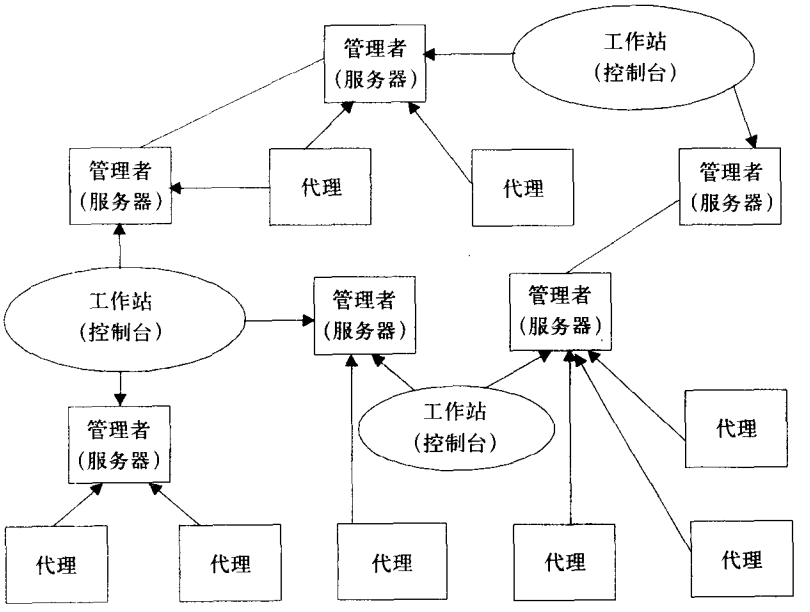
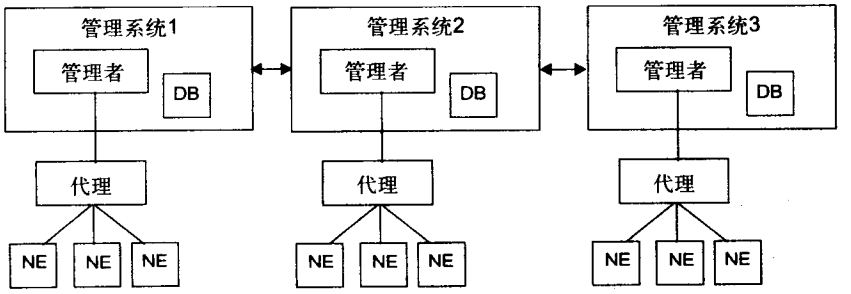


图22-17 基于多个管理者和工作站的分布式管理系统

如图22-17所示，每个代理控制一个特殊的网络元素，该网络元素的参数存储于对应的MIB中。

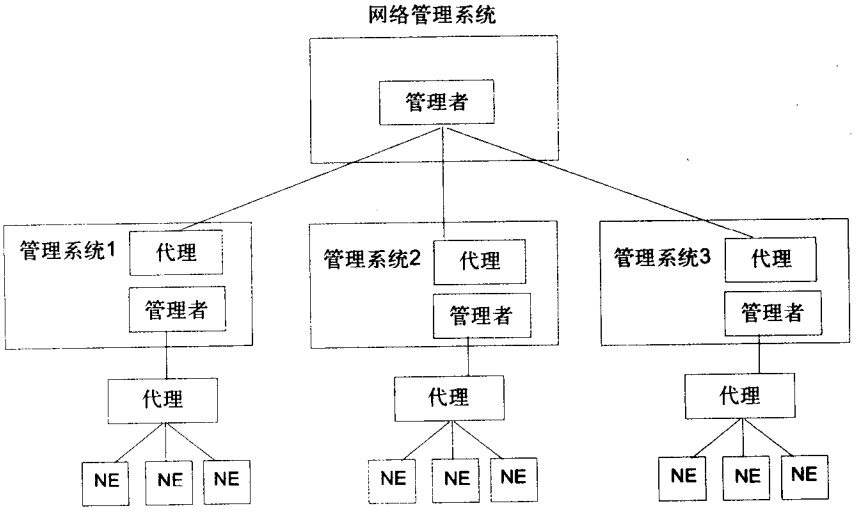
管理者们检索它们代理的MIB数据，对数据进行处理，然后把这些信息保存在特殊的数据库中。在工作站工作的操作员可以连接任何一个管理者，并使用GUI查看有关被管理网络的信息或发送控制指令给管理者以管理整个网络或网络的一些元素。

多个管理者的可用性允许负载在它们之间分配，因此保证了系统的可延拓性。通常，使用两种链路来连接管理者：点到点（图22-18）和层次（图22-19）。



DB —数据库
NE —网络元素

图22-18　管理者之间的点到点连接



NE—网络元素

图22-19　管理者之间的层次连接

在点到点连接中，每个管理者通过从下层代理那收到的信息控制自己的网段。这里没有中心管理者。管理者之间操作的协调是通过它们的数据库之间的信息交换来实现的。如今，点到点协议被认为是低效和过时的。

管理者之间的层次连接方法被认为是更强大更灵活的。每个低层管理者还扮演着它高层管理者的代理的角色。对于它的网段来说，这种代理有更一般化的更大的MIB模型。这个MIB积累了高层管理者为了控制整个网络所需要的信息。通常，不同层次网络模型的发展总是从顶层开始。在顶层，必须定义下面的管理者/代理所需的信息。这种自顶向下的方法减少了流通于网络管理系统不同层次之间的信息量。

管理者—代理—被控对象模型是如基于SNMP的因特网标准和基于**通用管理信息协议**(common management information protocol，CMIP)的ISO/OSI标准这些流行的管理标准的基础。

22.5.4 基于SNMP的管理系统标准

暂时的现象往往趋向成为永久。简单网络管理协议（SNMP）可以作为这个真理的又一证明。作为IP网络的临时的简单的解决方法而被开发的简单网络管理协议，现在变得非常受设备开发者和网络管理员的欢迎，以至于成为管理系统的头号协议，并占据此位很长一段时间。这是拥有ITU-T国际标准地位的管理协议CMIP都没能实现的，尽管它拥有更长的历史也更强大（相应的也更加复杂）。

然而，当更新版本的SNMP——SNMPv2出现时，却没有得到网络设备制造商的支持，也没有被广泛使用。来自IETF的开发者试图通过提供第三版本SNMPv3的规范来改善这个情况。这个协议的重大改进：保证灵活的网络管理系统的代理管理，控制信息的保护和对基于SNMPv1的系统的向后兼容性，及开放的体系结构允许SNMPv3的作者们可以期望它的成功。

虽然SNMP有提供给其他栈的实现，如IPX/SPX，但它却是一个为TCP/IP栈开发的应用层协议。SNMP用来从网络设备那获得有关它们状态、性能及其他存储于MIB的特性的信息。SNMP的简单性在许多方面是MIB SNMP数据库的简单性导致的，尤其是它们的最初版本，MIB-I和MIB-II。

在基于SNMP的网络管理系统中，下列元素是标准化的：

- 代理和管理者之间相互作用的协议，即SNMP自身。
- MIB模型和SNMP信息的描述语言，这是ASN.1抽象语法符号语言（ISO 8824：1987标准、ITU-TX.208建议）。
- 注册在ISO标准树上的一些MIB模型（MIB-I、MIB-II、RMON和RMON 2）的名字。这些标准定义，MIB结构，包括数据库对象类型、它们的名字和允许的对它们的操作（如读操作）。MIB的树结构包括强制（标准）子树和私有子树，允许智能设备的制造商根据特殊的MIB对象控制特殊设备的功能。

所有其他功能由管理系统开发者的意愿而实现。

SNMP是请求-响应协议，这意味着代理必须响应来自管理者的每个查询，这个协议的简单性是它的显著特征，因为它仅包括少量的指令：

- *GetRequest*——管理者根据对象名字用这个指令从对象处获得某个值。
- *GetNextRequest*——当连续观察对象表时，管理者用这个指令来检索下一个对象（没有指定它的名字）的值。
- *GetResponse*——SNMP代理使用这个指令将传递给管理者的指令回复给GetRequest或GetNextRequest指令。
- *Set*——这个指令允许管理者改变特殊对象的值。事实上，Set指令是执行设备管理的指令。代理必须正确解释用于控制设备对象的值。基于这些值，它必须执行管理操作，如阻塞一个端口或把一个端口赋给一个特殊的VLAN。Set指令同样适合于设定某个条件，在该条件下SNMP代理必须发送一条对应的报文给管理者。它也可能指定对一些操作的反应，如代理初始化、代理重启、连接终止、连接恢复、错误识别或下一路由丢失。如果发生了这些事件中的任何一个，那么代理将发起一个中断。
- *Trap*——代理用这个指令来通知管理者异常信息。

22.5.5 SNMP MIB结构

目前，SNMP使用多种MIB标准。几个主要标准为：MIB-I与MIB-II，及为远程管理设计的远程监控（RMON）MIB数据库版本。除此之外，还有为特殊MIB设备（如，传感器MIB或调制解调器MIB）及为某些设备制造商所有的MIB而设计的专用标准。

最初的规范，MIB-I，仅定义了读对象值的操作。像设置或改变对象值这样的操作是MIB-II规

范的内容。

MIB-I（RFC1156）定义了被分为8组的114个对象：

- *System*——关于某个设备的一般信息（如生产商的ID和上一次系统初始化时间）
- *Interfaces*——设备的网络接口的参数（例如，它们的编号、类型、交换速率和最小分组大小）
- *Address Translation Table*——对网络地址到物理地址的映射的描述（例如，依据ARP）
- *Internet Protocol*——有关IP的数据（如IP网关、主机的地址和有关IP分组的统计数字）
- *ICMP*——与ICMP相关的数据
- *TCP*——与TCP相关的数据（例如，有关TCP连接的信息）
- *UDP*——与UDP相关的数据（如传送的、接收的和损坏的UDP数据报的数量）
- *EGP*——与因特网中EGP操作相关的数据（如正确接收的报文数量和接收的有错报文的数量）

从这个清单中，可以明显看出MIB-I是严格地面向支持TCP/IP栈协议的路由器的。

在下一版本中，于1992年被采用的MIB-II（RFC1213），支持的对象清单被极大地扩充了（扩充为185个标准对象），并且组的数量也增加到10。

图22-20提供了MIB-II树结构的一个示例片断。它图释了两个（十个中的两个）可能的对象组——*System*（对象名字带Sys前缀）和*Interface*（名字带if前缀）。

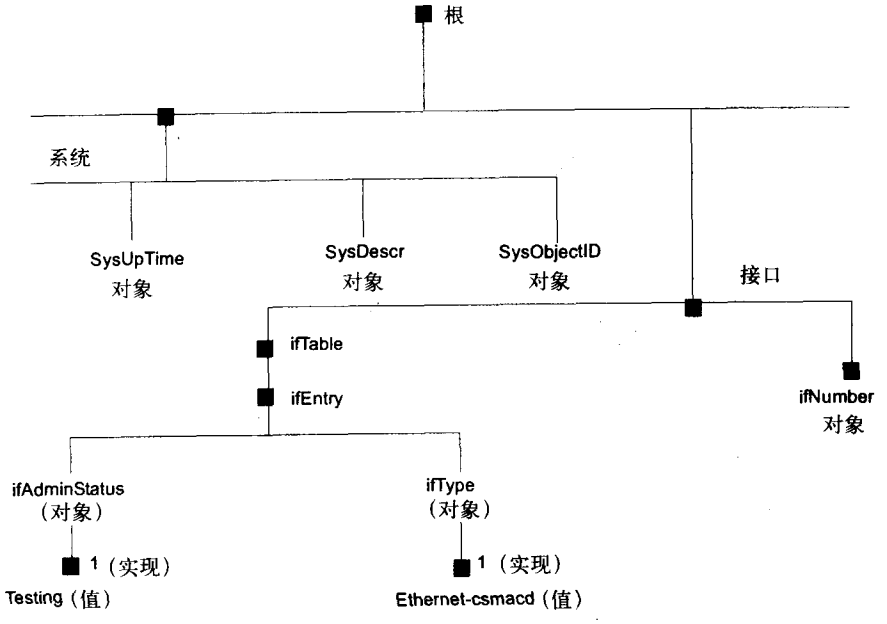


图22-20　标准MIB-II树的片断

对象SysUpTime包含系统距上次重启动时间，对象SysObjectID指定设备ID（例如，一个路由器的ID）。

对象ifNumber定义了设备网络接口数，对象ifEntry是描述某个设备一个接口的子树的顶点。对象ifType和ifAdminStatus是这个子树中分别定义设备类型和状态的部分（在这个例子中，这个接口为以太网的接口）。

描述每个具体设备接口的清单包括如下这些：

- ifType——该接口支持的协议类型。这个对象取值为所有的标准数据链路层协议，如rfc877-x25、Ethernet-csmac、iso88023-csmacd、iso88024-tokenBus及iso88025-tokenRing。

- ifMtu——可以通过该接口发送的网络层分组的最大大小。
- ifSpeed——接口带宽，单位为位每秒（快速以太网为100Mb/s）。
- ifPhysAddres——端口物理地址。对快速以太网来说，这将是MAC地址。
- ifAdminStatus——要求的端口状态。
- up——端口已准备好传送分组。
- down——端口没有准备好传送分组。
- testing——端口运行于测试模式。
- ifOperStatus——当前端口状态，这个值同ifAdminStatus。
- ifInOctets——自SNMP代理上次启动后，该端口收到的字节总数，包括控制字节。
- ifInUncatPkts——送往高层协议的带独立接口地址的分组序号。
- ifInNUcastPkts——送往高层协议的有广播或多播接口地址的分组数目。
- ifInDiscards——接口正确接收但没有送往高层协议的分组数目，原因可能是缓冲溢出。
- ifInErrors——因为探测到有错误，所以没有传递到高层协议的已发送分组数目。

除了描述流入分组的统计量的对象，这里还有类似的有关流出分组的对象。

正如从对MIB-II对象的描述中所看到的那样，这个数据库不提供有关以太网帧的典型错误的具体统计量。此外，它也不提供任何有关这些特性如何随时间变化的信息。注意到网络管理员对这个信息特别感兴趣。一段时间过后，这些限制被更新的MIB标准RMON MIB删除。RMON MIB是专门的面向以太网协议具体统计量收集的，它的能力包括为参数值建立时间关系。

为了命名MIB变量，同时无二义性地定义它们的格式，使用了一个附加的规范，叫做**管理信息结构**（structure of management information，SMI）。例如，SMI包括标准的IpAddress名字并定义它的格式为一个4字节的串。另一个例子是名字Counter，它是一个取值范围为0到$2^{32}-1$的整型数。

MIB变量的名字可以使用符号形式和数字形式来书写。符号形式用于在文本文件中表示变量并在屏幕上显示它们。在SNMP报文中使用数字名字。比如，符号名字SysDescr对应的数字名字为1.3.6.1.2.1.1。

SNMP MIB对象的数字组合对应于该对象在ISO对象注册树上被授予的全名。SNMP开发者没有使用传统的以太网标准，这些传统的以太网标准由一些作为注册参数的数值组成，这些数值被称为在特殊的RFC中被赋了值的数。他们已在如图22-21所示的ISO标准树中注册了MIB SNMP数据库的对象。

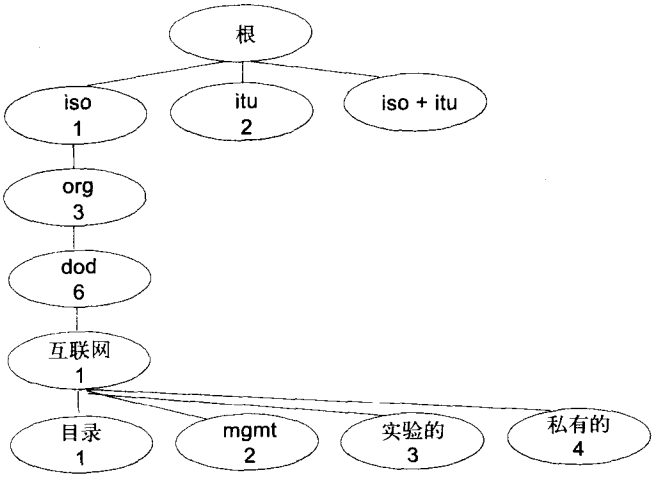


图22-21 ISO对象名字空间

像在任何复杂系统里一样，ISO对象的名字空间有一个树状层次结构。例子中图22-21仅展示了这个树的根。从这个根分出的三个分支分别对应ISO、ITU及两个组织联合体控制的标准。ISO为由国家和国际组织创建的标准建立了另一分支（这是org分支）。因特网标准是在美国国防部控制下开发的；所以，MIB标准适合标准的网络管理组下的dod-internet子树。任何在ISO下创建的标准的对象都由它们的自此树根往下的复杂的符号名字唯一标识。协议报文使用可

以唯一地映射到对应的符号名字的数字名字。名字树的每个分支都自左向右地用整数编号，这些编号代替了符号名字。因此，MIB对象的符号全名iso.org.dod.internet.mgmt.mib就被映射为数字名1.3.6.1.2.1。

对象的私有组（4）是为由Cisco和Hwelett-Packard这样的商业公司创建的标准预留的。CMIP和TMN类的对象命名使用同一棵注册树。

相应地，MIB-I和MIB-II对象的每个组，除了简短的符号名字外，还有完全规范符号名和对应的数字名字。

22.5.6 SNMP报文格式

SNMP在代理和管理者之间传送数据。SNMP使用的是不提供可靠报文传输的UDP数据报传输协议。在TCP基础上组织可靠的报文传输的该协议会使被管理设备负荷过重，它们在SNMP发展时代还不够强大。因此，SNMP开发者决定放弃使用TCP。

与大多数其他通信协议的报文相比，SNMP报文的报头都没有预先确定的字段。任何SNMP报文都由任意数目的字段组成，并且每个字段都由描述符事先指定它的类型和大小。

每个SNMP报文都由下面三个主要部分组成（图22-22）。

- 协议版本（版本）。
- 共同体标识符（共同体）用来把对象聚合到被某管理者控制的管理域。共同体标识符类似于口令。对于使用SNMP进行通信的设备而言，它们必须有和标识符匹配的值（通常默认使用public串）
- 数据域中包含前面描述过的协议指令、对象名字和它们的值。数据域包含一个或多个数据单元（PDU）。每个PDU对应下面的一个SNMP指令：

GetRequest-PDU，GetNextRequest-PDU，GetResponse-PDU，SetRequest-PDU，Trap-PDU。

*PDU*字段可能包含前面列举的四个指令中的一个。指令结构如图22-22所示，图示为GetRequest指令下的SNMP帧。

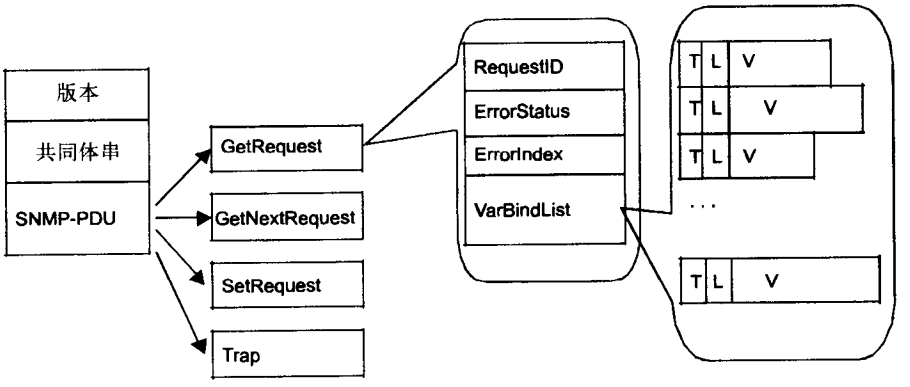


图22-22　SNMP 报文格式

变量RequestID是4字节的整型值（用于映射对请求的回复）。ErrorStatus和ErrorIndex是1字节的整型数，它们在请求中必须设为0。VarBindList值是某个管理者所需要的对象的数字名字的列表。这个列表包含类型、长度和值（T，L和V）这个三元组，它考虑了对任意数量的任意类型的变量的灵活规范。在请求期间，变量值必须设为null。

下面是SNMP报文的一个示例，它描述了对SysDescr对象值的请求（该对象的数字名字为1.3.6.1.2.1.1.1）。

30	29	02	01	00			
SEQUENCE	len=41	INTEGER	len=1	vers=0			
04	06	70	75	62	6C	69	63
string	len=6	p	u	B	l	i	C
A0	1C	02	04	05	AE	56	02
getreq	len=28	INTEGER	len=4	——————request ID——————			——
02	01	00	02	01	00		
INTEGER	len=1	status	INTEGER	len=1	error	index	
30	0E	30	0C	06	08		
SEQUENCE	len=14	SEQUENCE	len=12	Objectid	len=8		
2B	06	01	02	01	01	01	00
1.3	6	1	2	1	1	1	0
05	00						
null	len=0						

如前所述，报文从代码30开始（所有编码都是16进制），它对应于关键词SEQUENCE，一种支持的SNMP变量数据类型。该关键词把整个SNMP报文定义为一个变量序列。这个序列的长度在下一字节中指定（41字节），它对应于整个报文的长度。接在它后面的是一个1字节的整型数，它代表SNMP版本（在这个例子中，这个值被置为0，对应于SNMPv1；值1则对应于SNMPv2）。共同体字段为string类型。这个字符串长为6字节且可以取值为public。如果管理员没有指定其他值，则所有代理都默认支持该值。只有当默认值匹配时，代理才回复管理者的请求。

报文的另一部分由GetRequest指令组成。GetRequest操作由代码A0标识（这个值定义在SNMP中而不是ASN.1）。这个数据块的总长度为28字节。根据GetRequest块的结构，接在数据块后的是请求标识符（它被定义为一个4字节的整型数）。还有状态和和错误索引这两个1字节的整型数，它们在这个请求中被置为0。最后，报文由对象列表终止，它在这个例子中仅包括一个名字为1.3.6.1.2.1.1.1.0值被置为null的对象。

22.5.7 RMON MIB规范

改进了与MIB远程通信能力的RMON MIB规范，是SNMP的一个附加功能。RMON MIB数据库包含了关于设备的聚合信息，它不要求通过网络传送大量的信息。因为这个特征，远程控制变得很便利。RMON MIB对象包括附加的分组故障计数器，更灵活的收集统计量并分析趋势的工具，更强大的捕获并分析单独分组的过滤器及更复杂的警告信号设置环境。RMON MIB代理的智能比MIB-I或MIB-II代理的都要高，这使得它们可以执行大多数处理设备信息的工作。在RMON MIB出现之前，这些任务通常都是委托给管理者的。这些代理可以嵌入到通信设备中或作为运行于普通PC机或笔记本电脑上的独立软件模块。

在MIB对象集合中， RMON对象被赋予编号16。RMON对象由十组对象组成（列表中省略的第十组由令牌环协议的特殊对象组成）：

- *Statistics*（1）——当前积累的分组特征统计量，例如冲突次数等。
- *History*（2）——为将来的趋势分析而和预定义的周期一起保存的统计数据。
- *Alarms*（3）——统计参数的阈值。当超过这些值时，RMON代理发送一条报文给管理者。

- *Hosts*（4）——网络主机上的数据，包括它们的MAC地址。
- *Host TopN*（5）——负载最高的网络主机列表。
- *Traffic Matrix*（6）——网络内每对主机间流量强度的统计数字，以矩阵形式排序。
- *Filter*（7）——分组过滤条件。
- *Packet Capture*（8）——分组捕获条件。
- *Event*（9）——事件产生及记录的条件。

这些组以它们列出的顺序编号；例如Hosts组，数字名字为1.3.6.1.2.1.16.4。

RMON MIB标准定义的对象总数约为200。它们分为10组，记录在两个文件中——用于以太网的RFC1271和用于令牌环的RFC1513。

RMON MIB的特征是它与网络层协议的独立（与面向TCP/IPMIB-I和MIB-II标准相比）。所以，它便利于使用多种网络协议的异构网络。

考虑一下Statistics组的细节，它定义了RMON代理可以提供的以太网帧（标准叫法为分组）的信息。History组是基于Statistics组对象之上的，这是因为它的对象允许为Staticstis组对象建立时间关系。

Statistics组，连同其他的一些对象，有以下这些：

- etherStatsDropEvent——当分组因为缺少可用资源而被代理忽略时的事件总数。接口无需丢弃这些分组。
- etherStatOctets——从网络接收的字节（不包括前导码但包括校验和）总数（包括错误分组）。
- etherStactsPkets——接收到的分组总数（包括错误分组）。
- ehterStatsBroadcastPkts——发送到一个广播地址的完全无错分组的总数。
- etherStatsMulticasttPkts——一个多播地址接收到的完全无错分组的总数。
- etherStatsCRCAlignErrors——接收到的长度（不包括前导码）范围为64到1 518字节且字节数不是整数（队列错误）或有校验和错误（FCS错误）的分组的总数。
- etherStatsUndersizePkts——长度小于64字节但被正确格式化的分组总数。
- ehterStatsOversizePkts——长度超过1 518字节但被正确格式化的分组总数。
- etherStatsFragments——字节数不是整数或有校验和错误且大小不足（长度小于64字节）的分组总数。
- ehterStatsJabbers——字节数不是整数或有校验和错误且长度大于1 518字节的分组总数。
- etherStatsCollisions——对给定的以太网网段内冲突数目的最好的评估。
- etherStatsPkts64Octets——接收到的大小64字节的分组（包括损坏的）的总数。
- etherStatsPkts65to127Octets——接收到的长度为65到127字节的分组（包括损坏的）的总数。
- etherStatsPkts128to255Octets——接收到的大小为128到255字节的分组（包括损坏的）的总数。
- etherStatsPkts256to511Octets——接收到的长度为256到511字节的分组（包括损坏的）的总数。
- etherStatsPkts512to1023Octets——接收到的大小为512到1 023字节的分组（包括损坏的）的总数。
- etherStatsPkts1024to1518Octets——接收到的大小为1 024到1 518字节的分组（包括损坏的）的总数。

从对对象的描述中可以看出，使用嵌入到转发器或其他通信设备中的RMON代理，就有可能执行对以太网或快速以太网网段操作的详尽分析。首先，有可能获得有关帧错误的数据，尤其是这个网段；其次，使用History组对这些错误建立的时间关系会非常有用（也有可能把它们绑定到某个具体时刻）。获得时间关系分析的结果后，就可能得出错误帧可能来源的初步结论。基于这个信息，管理员就能提炼出有更具体属性（通过使用Filter组对象指定它们）的帧的捕获条件，这相

当于研究造成错误的版本。在那之后，通过分析捕获的帧就有可能执行更详尽的分析，这些帧是从Packet Capture组的对象那获得的。

后来，RMON 2标准被采用。它把RMON MIN数据库智能的想法扩展到更高层协议，并执行部分协议分析器的功能。

小结

- IP WAN可以分为两类：纯IP网络和覆盖IP网络。在覆盖网络里，最流行的是在ATM上的IP、在帧中继上的IP和在MPLS上的IP。
- 在纯IP网络中，没有其他层是在分组交换技术基础上运行的。在这种网络里，因为选择最短路由的必要性而导致的路由器上的流量路径的不确定性和负载分配效率低下，使得保证QoS和解决流量工程任务变得很复杂。
- 在纯IP网络的数据链路层上，使用的是点到点协议，它确保相邻路由器之间的帧传输。如今，这类协议中最受欢迎和使用最广泛的是HDLC和PPP。
- HDLC是最老的协议，它是为可靠地通过噪声信道的数据传输而于20世纪70年代开发的。它使用滑动窗口算法来保证对帧传输的控制和对丢失或损坏帧的恢复。
- PPP是为当代高质量的通信链路而开发的。这个协议不解决有关可靠传输的问题，但它允许在建立连接时用一个灵活的协商程序对相邻设备的参数进行协调。除此之外，它还确保设备相互识别。
- 在如ATM上的IP或帧中继上的IP这样的覆盖网络中，路由器是通过虚拟ATM或帧中继信道而不是物理链路连接的。这提供了使用流量工程确保合理网络负载和使用ATM服务系统来为每个流量等级确保一个具体的QoS级别的可能性。
- 新的MPLS技术是和IP栈技术紧密结合的。结合了IP路由器和MPLS交换机的功能的设备渐渐以LSR而闻名。
- LSR不但使用TCP/IP栈路由协议来选择合理的虚拟路径LSP，还用它来检测网络拓扑和状态。
- MPLS可以使用不同的数据链路层技术的帧，如PPP、以太网、ATM和帧中继。
- MPLS有三个应用域，相应地也有三种变体：MPLS IGP、MPLS TE 和MPLS VPN。
- MPLS IGP使用LDP信令协议，并自动建立到网络路由器知晓的所有目的网络的路径。 然而，这些路径是为TCP/IP栈的IGP标准所选取的，所以，相应地也不解决流量工程问题。
- MPLS TE基于修改的IS-IS或OSPF路由协议运作，它不但传播包含拓扑的信息还传播包含关于可用链路带宽信息的路由广告。这允许建立TE LSPs，即隧道，为聚合数据流保留带宽。
- 网络管理系统的功能在ITU-T X.700建议和ISO 7498-4文件中被标准化。它有5个功能组：配置管理，故障管理，性能管理，安全管理和计费管理。
- 网络管理系统的基础是管理者—代理方法。这个方法使用一个叫做MIB的被管资源的抽象模型。
- 代理使用一个定制接口与被管资源通信。代理和管理者之间的通信是使用一个标准协议通过网络执行的。IP网络使用了SNMP。
- 因特网标准的第一批MIB数据库是面向路由管理的：MIB-I仅为了管理，MIB-II既为管理又为了控制。RMON MIB是它们的更进一步发展，它目标在于控制更低层以太网和令牌环的接口的智能代理的开发。因特网MIB数据库的标准对象的名字被注册在ISO标准名字树上。

复习题

1. 是什么促进了一些IP WAN（纯IP，ATM上IP，帧中继上IP和MPLS上IP）模型的发展？
2. ATM上IP或帧中继上IP网络由两层分组交换网络组成，而MPLS上IP网络仅由一层组成，这种

说法对吗？说明你的理由。

3. 比较HDLC和PPP的主要特征。它们中的哪些是优点，哪些是缺点，在哪些条件下？
4. 在HDLC协议和PPP协议里连接建立程序的目的是什么？
5. HDLC使用什么机制来恢复丢失或损坏的分组？
6. PPP多协议的功能有哪些？
7. 为什么需要PPP提供的相互识别程序？
8. 列举为了使用租用线路操作路由器，配置程序配置的主要阶段。
9. 在IP-over-ATM模型中ATM层的作用是什么？
10. IP交换技术实现了哪些新想法？
11. MPLS保留了哪些IP交换的想法，修改了哪些？
12. 列举在LSR中，IP路由器的主要功能模型。
13. 使用MPLS标记栈可以确保哪些新的可能性？
14. 假设LSR使用以太网帧格式。这是否意味着设备会基于在IEEE802.1D标准基础上获得的路由表转发帧？
15. 如何建立一个经过多个MPLS域的LSP？
16. MPLS IGP和MPLS TE的区别是什么？
17. 在MPLS IGP操作中有必要手动配置LSR吗？
18. 在ATM和帧中继技术中与MPLS TE隧道相似的技术是什么？
19. 在支持MPLS的网络中，是否有可能使用一般的IP转发技术传递部分流量？
20. 根据X.700标准列举网络管理系统的功能组。
21. 网络管理系统和系统管理系统之间有区别吗？如果有，那么它们的区别是什么？
22. 在网络管理系统中被委托给代理的任务是什么？管理者执行的功能是什么？
23. 列举标准MIB 。
24. 在SNMP中，MIB对象使用的是那种名字——符号的还是数字的？
25. 什么时候使用Trap指令？

练习题

1. 假设你是一个IP WAN网络设计者。你会向想要你开发一个大型IP网络的客户提出什么问题，以便你能选择多层模型的类型（非覆盖IP，ATM上IP，帧中继上IP或MPLS上IP）？
2. 测量显示某个通信链路的BER值为10^{74}。你将为该链路选择哪种协议——HDLC或PPP？
3. 为图22-13所示的LSR1构造一个转发表。
4. 为了使图22-14所示的某个网络管理员能够解决流量工程问题，必须收集网络哪些数据？ 使用你的数据版本并解决该问题。

第23章 远程访问

23.1 引言

当一个家庭用户计算机需要访问因特网或位于LAN覆盖范围外的公司的私有网络时，经常会使用到“远程访问”这个术语。在这种情况下，就需要使用WAN链路。最近，远程访问的概念不但包括独立的家庭计算机的网络访问，而且已经开始包括连接多个家庭用户计算机的家庭网络的网络访问。总共只有两到三个雇员的小公司的办公室也有这种小网络。

远程访问的组织是当前计算机网络最急需解决的问题之一。它也已成为有名的“最后一英里问题”。在这个例子里，“最后一英里”是最近的通信运营商汇接点（POP）与用户建筑物的距离。解决这个问题的难处有多个方面。用户需要保证所有类型通信量的高质量传输的快速访问，这些通信量包括数据、声音和视频。为了这个目的，就必须支持每秒几兆比特或至少每秒几十万比特的传输速率。然而，城市里的绝大多数建筑，仍是由本地电话环路连接到通信运营商的POP，这些环路把用户的速率限制在每秒几万比特。这个问题在农村地区尤其严重。

在不久的将来，电缆基础设施的大规模重建几乎是不可能的，因为这个任务实在是太大了。考虑到分布在广阔空间上的建筑和房屋的数量。虽然在某些工业国家，通信运营商已经开始在建筑物和私宅里安装高速光纤链路，但是这样的国家毕竟是少数。而且，目前只有那些有大量的潜在用户的大型城市和大型建筑物才涉及这个过程。

长期以来拨号访问一直是最流行的远程访问技术。用户通过使用这种方法建立调制解调器到因特网或者公司网络的拨号连接。这种方法有一个重大缺陷：访问速率被限制在每秒几万比特以内。这个限制是必然的，因为每个电话网络的用户都被赋予了一个约为3.4KHz的固定带宽（回想我们在第9章提到的电话网络中使用的多路复用技术）。越来越多的用户对此访问速度不满。

目前，有多种组织高速远程访问的新技术，它们都使用已有的本地电话环路或有线电视环路。在到达服务提供商的POP后，计算机数据不是经由电话或有线电视（CATV）网络来传送的，而是用特殊设备传递到数据传输网络上的。这提高了访问速率，并克服了分配给每个电话或CATV网络用户的隐含的带宽限制。这类技术中最受欢迎的是非对称数字用户线（ADSL），它使用本地电话环路和运行于CATV网络之上的电缆调制解调器。这些技术确保了一个从每秒几十万比特到每秒几兆比特的通信量的速率。

除此之外，还使用了多种无线访问技术来保证固定或移动访问。使用的无线技术集合除了多个固定访问技术，如基于新802.16标准的那些外，还包括无线以太网（802.11）、多种专利技术、通过移动电话网的传输。

在本章中，我们将提到了远程访问的几种最流行的方法和技术。

23.2 远程访问的方法

图23-1描绘了混合且多样的远程访问世界。这里你可以看到多种类型的客户，他们的区别在于使用的设备和对访问参数的要求。此外，还有多种连接客户建筑与最近的通信运营商POP（或电信提供商的中心办事处）的方法。例如，可以使用一个本地模拟的或数字的环路、TV电缆或无线链路来实现。最终，通信运营商可能擅长于不同类型的服务，这意味着它可能是一个电信服务提供商、一个ISP、一个CATV提供商或一个普通的提供整套服务并拥有所有类型网络的提供商。

23.2.1　客户和终端设备的类型

仔细考虑图23-1所示的访问方法的每个元素。

客户1和客户2是最典型的用户，因为他们两个都只有一台需要提供远程计算机网络访问的计算机。除了计算机外，这些用户还使用电话和电视机（TV），所以，这些设备的终端设备可以用来组织这台计算机到数据传输网络的远程访问。

客户2使用两个本地电缆环路，第一个类似于一个基于双绞线的模拟的电话本地环路，第二个为使用同轴TV电缆的CATV本地环路。这些本地环路有不同的特性。如，客户建筑物与提供商POP之间的长度为1～2km的双绞线，带宽通常为几兆赫，而同轴电缆则保证有几十兆赫的带宽。

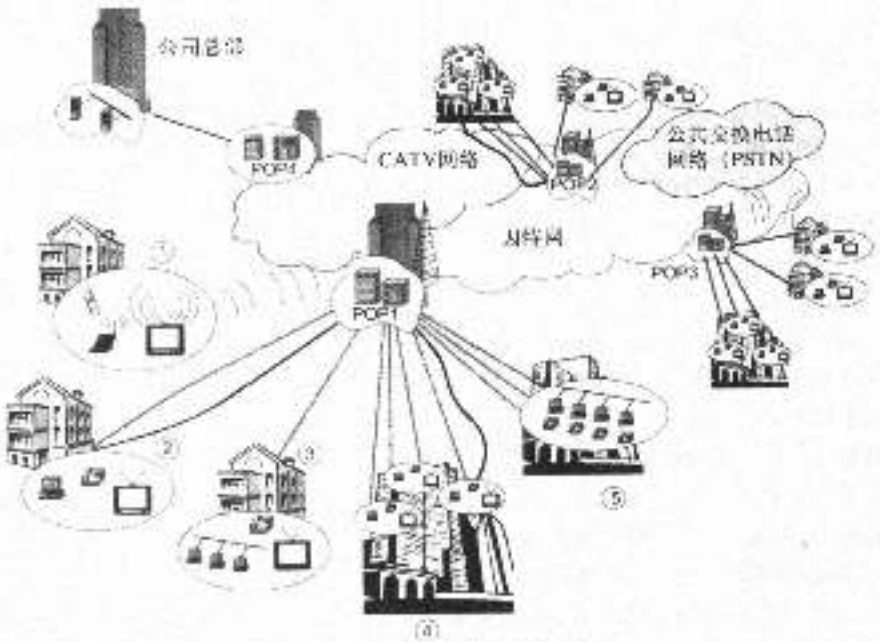


图23-1　远程访问客户

客户1没有有线本地环路，因为这个用户使用的是一个移动电话。这个用户也不使用有线电视服务，他通过无线链路接收TV信号。

因此，为了给客户2组织远程访问，提供商既可以使用已有的TV电缆也可以使用电话本地环路。对客户1来说，没有这种可能性；所以，对这个客户，提供商必须做到要么保证无线通信，要么安装一条连接这个客户建筑物与最近POP的新电缆。

客户1和客户2的最典型特征是非对称型的通信量，因为家庭用户通常从因特网下载信息。像ADSL这样的非对称技术是最适合这种要求的。

客户3与客户1和客户2不同，因为它有多台连接到某个LAN的计算机。个人和小公司都可能是这种类型的用户。对于LAN而言，保证其远程访问的特征就是提高带宽要求。此外，如果家庭LAN包含一台给因特网用户或公司的分公司雇员提供信息的服务器，则这种客户的通信量可以是对称的。因为客户3没有CATV本地环路，所以只有电话本地环路可以用来为它提供远程访问。客户3可以使用多种方法组织自己的IP网络。例如，它可以向提供商请求一个IP地址池，这样它的网络中的每台计算机都将有一个永久的公共IP地址。对这个客户而言，这是最灵活的变体，因为客户网络的每台计算机都可以是因特网的一个正式成员。它们不仅扮演着客户机器的角色而且还扮演着有注册域名的服务器的角色。在这种情况下，客户LAN必须有一个边界路由器。客户LAN可

以通过它与提供商网络通信。IP网络的另一种变体是基于第20章所述的NAT技术。

客户4由公寓大楼的住户组成。公寓大楼使用CATV电缆和电话本地环路的多路双绞线（每间公寓一个本地环路）连接到提供商POP。仅使用一条CATV电缆为大量客户服务会产生额外的问题，因为电缆起着一个共享介质的作用。使用本地电话环路为公寓的住户组织远程访问与连接一个用户（如客户2）没有区别。虽然大多数这样的用户使用的是传统的模拟本地环路，但是在这座公寓大楼里还是有很多公寓的房客是综合业务数字网（ISDN）服务的用户。这意味着虽然他们的本地环路是基于双绞线的，类似于传统的模拟环路，但是这些其实是数字的本地环路。虽然ISDN最初是作为一个为数据传输服务提供电话服务的通用网络而设计的，但是，实际上它被用作一个普通的电话网络。

客户5也是由公寓大楼的住户组成。然而，在这种情况下，提供商在公寓大楼里安装了一个LAN。所有决定订阅该提供商提供服务的客户都连接到这个LAN。对提供商来说，只有在潜在的客户足够多时这个变体才是有效率的。安装于一栋公寓大楼里的LAN比个人计算机或个别客户的家庭网络需要更高的访问速率。所以，提供商必须保证用于组织远程访问的本地环路有宽阔的带宽。为了这个目的，可能会要使用已有的CATV电缆、一个专门安装的以太网电缆、甚至一个光纤电缆。

远程访问提供商要么服务所有类型的客户，要么专攻于某种服务，如给私宅或小企业的所有者提供服务。一个通用的提供商必须支持组织最后一英里的任何变体，这不可避免地使需要的设备和访问技术复杂化。

对任何本地环路，提供商必须保证通过此本地环路的位传输，并把它和该本地环路最初设计的信息传输组合起来，如声音或CATV信息。然后基于该物理层设备，提供商必须给客户提供请求类型的访问服务。

访问提供商必须解决的一个问题是，为通过物理的方式连接到其他通信运营商的本地环路的客户组织访问。使用图23-1所示的配置为例。这里，提供商A拥有POP1和POP2，提供商B拥有POP3。如果提供商A想给连接到POP3的客户提供网络访问服务，那么提供商A必须和提供商B达成一个适当的协议。这个协议可能管理提供商之间通信的多种方法，这在第5章都已说明。例如，提供商A可能从提供商B那里租用它的客户所使用的那些本地环路，用这些环路来转发从它们那里收到的数据到提供商A自己的网络，然后再根据用户的需要转发这条信息。另一种情况是，本地环路可能仍然由提供商B支配，但提供商B必须把计算机数据从电话和TV信息中分离出来，并把它们转发到提供商A的网络中。在两种情况下，都需要保证提供商A和提供商B的网络之间的通信。

最简单的因特网访问变体是提供客户与公司网络服务器的**无保护连接**（unprotected connection），这可能是不安全的。首先，在因特网上传输的机密数据可能被窃听或修改；第二，当使用这种方法时，对公司网络的管理员来说，很难限制未授权访问。这是因为事先并不知道合法用户（公司或企业的雇员）的IP地址。所以，大多数企业都偏爱基于虚拟专用网络技术的安全访问。这个技术将在下一章详细讨论。

23.2.2 在本地环路的信息多路复用

如图23-1所示，大多数私人建筑和公寓建筑都是通过电话或CATV本地环路连接到POP的。

所以，为了给客户提供如今最常用的三种访问方式（电话访问、TV访问和因特网访问），就必须保证多种类型数据在同一通信链路上的同时传输。例如，可能需要把使用相同电话本地环路的声音和数据传输组合起来，或把经由同一同轴电缆的数据传输和电话信号传输组合起来。

我们需要一种能同时传输所有三种类型信息的本地环路。不幸的是，双绞线不能胜任这种工作，因为相隔几千米它的带宽就只有几兆赫，通常是1 MHz。很明显，对每秒几兆赫的声音、视频和计算机数据的同时传输来说，这是不够的。

所以，只有同轴CATV电缆和宽带无线链路可以用作稳固的本地环路。自然地，我们会提到已有的和广泛使用的各种本地环路。当涉及给大型的新建筑物安装新电缆时，通常光纤电缆就会被添加到这个清单上。

实际上，将在后面几节提到的所有的访问技术都使用本地环路上的两种。有时全部的三种类型的信息多路复用。因此，ADSL使用模拟电话本地环路多路复用声音和计算机数据，同时电缆调制解调器组合通过同轴电缆的TV信号和计算机数据。多种无线访问技术保证了在同一本地环路内TV信号和计算机数据的传输，有时还有电话。唯一的例外是最早的技术，拨号访问。在这种情况下，模拟本地环路由电话和连接到某台计算机的拨号调制解调器交替使用。

使用通用本地环路组织访问的方法如图23-2所示。



图23-2　在本地环路中三种信息的多路复用

在本地环路中，信息多路复用使用最多的是频分多路复用（FDM）。根据用户的需要，每种类型的信息都被分配了明确的带宽。对电话连接，分配的带宽是4KHz，这对应于分配给模拟电话网络用户的标准带宽。对于计算机数据，则需要更宽的带宽。在非对称访问的情况下，必须为主要的下行通信量分配至少几十万赫兹的带宽，几百兆赫则更好。强度低些的上行通信量则要求有几万赫兹的带宽。有线电视通常为每个用户使用6MHz的带宽；然而，在这种情况下，只传送下行通信量。

为了实现选择的FDM方法，在用户建筑物和POP中安装有分路器（splitter）。分路器执行信号的多路复用和解多路复用。大多数情况下，分路器是一个把需要的频带分离出来并把每个频带传送到单独的输出端的无源过滤器。用户的终端设备，如电话、TV或计算机，都被连接到分路器的输出端。因为计算机使用离散信号进行数据交换，所以它需要一个把离散信号转换为所需频带的模拟信号的附加设备。

大多数用户习惯于使用拨号调制解调器（dial-up modem）。它使用模拟电话网络的标准的4KHz带宽。拨号调制解调器不与其他设备共享带宽，它的带宽全部用于计算机数据传输。在这种情况下，没必要安装分路器。

除了拨号调制解调器外，还有**ADSL和电缆调制解调器**（cable modem），它们分别运行于电话本地环路和CATV电缆。在这种情况下，要求有分路器，因为在这些本地环路上其他信息是和计算机数据一起传输的；电话或TV信号是这些本地环路的主要信息类型。

在提供商的POP上，每个本地环路都被连接到一个分路器，这个分路器在连接的另一端执行类似的多路复用和解多路复用。结果，电话信息由分路器的电话输出端提供给提供商的电话交换机，并由它把电话信息传送到电话网络。TV信号从适当的分路器输出端发送给连接到该提供商CATV网络的CATV设备。

最后，计算机数据被提供给收集计算机通信量然后把通信量传递给提供商LAN的设备。这个设备有多个名字。在图23-2中，它被标以最流行的一个名字，**远程访问服务器**（remote access server，RAS）。有时可能遇到这个设备的其他名称，例如，**远程访问集中器**（remote access concentrator）、**访问多路复用器**（access multiplexer）或**终端系统**（termination system）。简单起见，我们将使用最流行的术语，RAS。不管这个设备叫什么，所有这种设备都有相同的结构。它们包含大量的执行与用户调制解调器相反操作的调制解调器；即，它们调制下行通信量并解调上行通信量。除调制解调器外，RAS还包含一个从调制解调器收集通信量并把它传送给POP LAN的路由器。通信量从这个LAN被传送给因特网或以常规方法传送给某个公司网络。

我们已经描述了一个通用的访问方法，它依赖于选择的本地环路和调制解调器类型，产生多种访问技术。需要强调的是，就OSI模型而言，所有这些技术都是物理层技术，因为它们在客户计算机和提供商LAN之间创建了位流。为了让IP运行于该物理层之上，就必须使用一个数据链路层协议。目前，组织远程访问使用最多的是PPP，因为它支持诸如给客户计算机分配IP地址及用户识别这些重要的功能。

23.2.3 远程节点方式

如今，最流行的远程访问服务是**提供到公共因特网域的访问**（providing access to the public Internet domain）。这个服务假定提供商保证IP通信量的路由，这个通信量存在于用户计算机和任何有公共地址（或一个为公共访问使用NAT技术的私有地址）的因特网网站之间。

通常，提供商使用**远程节点方式**（remote node mode）来保证独立计算机的用户的访问。这种方式允许客户计算机加入某个远程LAN，并允许它访问物理连接到那个LAN的端节点用户可以使用的整个范围的服务。

出于这个目的，提供商通常保留一个可以赋给它的某个子网的RAS客户的IP地址池。对不需要永久因特网访问的客户，这个服务是作为一个拨号服务提供的，并且IP地址只是在连接时才动态分配的。为了节约子网地址，必然使用远程节点方式，因为在标准方式IP路由器必须给它的每个端口赋予一个不同的IP子网地址。显然，对于大多数客户网络的单个节点来说，这过于冗余了。对需要一个永久连接的客户来说，IP地址既可以静态分配也可以动态分配。

为了保证远程节点方式，提供商的RAS支持第17章提到过的代理ARP。这个特征，见图23-3，使得RAS有别于普通的IP路由器。

对属于地址为200.25.10.0/24的提供商的LAN的远程节点，网络管理员已分配了范围为从200.25.10.5到200.25.10.254的地址池。如果客户使用拨号服务来连接提供商网络（例如，使用PPP），它将从该地址池中临时获取一个地址。如此，客户1的计算机被赋予地址200.25.10.5，客户2的计算机被赋予地址200.25.10.6。当这些远程节点连接到网络时，远程访问服务器把下列记录插入到一张专门的类似于ARP表的表中：

200.25.10.5——MAC——P1

200.25.10.6——MAC——P2

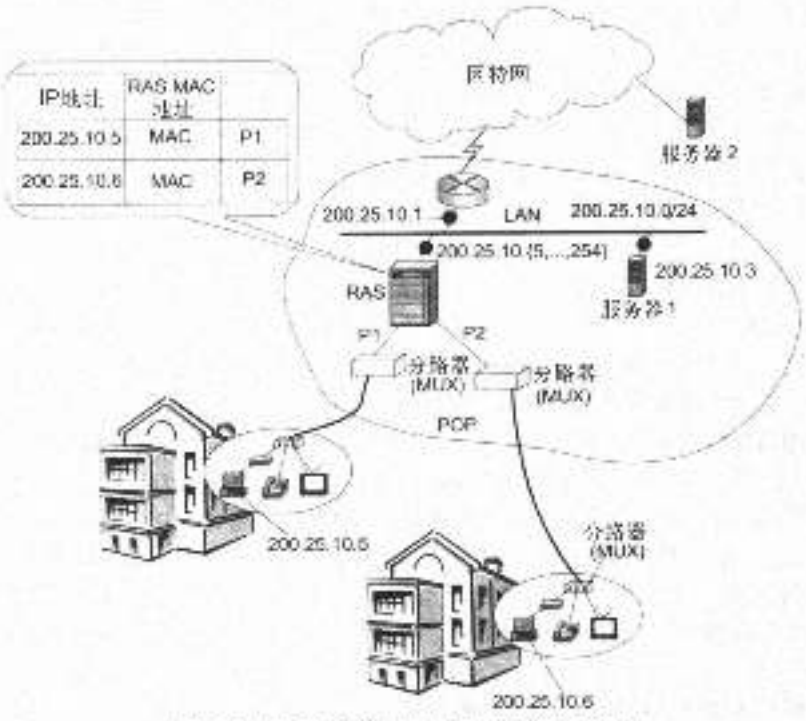


图23-3 组织远程访问时使用ARP代理服务器

MAC在这里指明远程访问服务器的内部接口的地址，P1和P2是RAS客户所连接的端口编号。

因此，端口编号在ARP表中扮演着MAC地址的角色。例如，如果连接到其中一个ISP的服务器2发送一个分组给客户1的计算机，那么该ISP的路由器R1则认为该分组是直接发送给直接连接到子网200.25.10.0/24的某个节点的。所以，R1发送一个包含地址200.25.10.5的ARP请求。RAS将代替客户1的计算机回复这个请求，把它的MAC地址提供给路由器R1。此后，R1发送带有该RAS的MAC地址的封装为以太网帧的IP分组，然后，RAS从发送给它的以太网帧中检索该IP分组，接着，基于这个IP地址从表中找出该分组需发往的端口编号。在这个例子中，该端口为P1。然后，RAS把该分组封装入PPP帧中，该帧用于对连接RAS与客户1的计算机的本地环路进行操作。

当客户拥有自己的LAN时，该LAN的主机已注册有公共的IP地址，RAS就作为一个普通的路由器运行。在这种情况下，这种运行模式就不叫远程节点方式了。

23.2.4 远程控制方式 Telnet

终端访问（terminal access），也叫**远程控制**（remote control），是一种特殊的访问方式。这种方式假定用户把他的或她的计算机变为他或她远程访问的另一台计算机的终端。

在计算机网络形成期间（如，在20世纪70年代），支持这种方式曾是网络的一个主要功能。X.25网络的PAD的存在，就是为了保证居住于其他城市和工作在最简单的字符终端的用户对主机的远程访问。

远程控制方式由一个专门的应用层协议来保证的，该协议运行于那些保证远程节点到计算机网络的传输连接的协议之上。目前有很多远程控制协议，既有标准协议也有专用协议。在IP网络中，最早用于这个目标的协议是telnet（RFC854）。

Telnet限制用户只能使用命令行方式，保证字符终端的仿真。Telnet根据客户端——服务器架构运作。

当用户在键盘上按下某一个键时，telnet客户端截取它的扫描码，把它封装到TCP报文中，然后通过网络把它发送到用户想要控制的节点。当分组到达目的节点时，telnet服务器从TCP报文中获取按键的扫描码，并把它传递给运行于该节点上的操作系统。操作系统把一个telnet会话当作是一个本地用户打开的会话。如果用户按某键，然后操作系统要输出对应的字符到显示器上作为响应，那么该远程控制会话将把这个字符封装到一个TCP报文中，然后通过网络把它传回给远程节点。Telnet客户端获取该字符，并在终端仿真窗口把它显示出来。

Telnet曾是为UNIX环境实现的，它连同电子邮件（e-mail）和通过FTP对文件文档的访问，曾是最受欢迎的因特网服务之一。如今，这个协议已经很少用于公共因特网域，因为没有人愿意第三方控制他的计算机。虽然telnet使用密码来防止未授权访问，但是这些密码是以明文形式在网络上传送的。因此，它们很容易被窃听并用于未授权访问。所以telnet的主要使用范围局限于某个单一的LAN，在那里密码窃听的概率要低很多。目前，telnet广泛用于控制通信设备，如路由器、交换机和集线器，实际上它已经不再用于控制计算机了。因此，它不用作用户层面协议。它变成了一个管理层面协议，作为SNMP的另一选择。

Telnet和SNMP的区别在于，telnet需要有一个人类管理员参与网络管理进程。这是因为配置或监控路由器或任何其他通信设备的管理员，要手动发出需要的指令，然后由telnet传送这些指令。与telnet相比，SNMP被设计为自动监控和控制，虽然它不排除管理员参与这个过程的可能性。为了排除在网络上明文传送密码导致的潜在危险，通信设备加强了它们的保护级别。通常，当密码以明文方式通过网络传送时，使用多级访问方法，仅提供读通信设备配置的基本特征的可能性。允许改变设备配置的管理访问则需要使用另一个以加密形式传送的密码。

远程控制也可以以GUI模式执行。对马萨诸塞州技术协会开发的UNIX来说，X视窗系统是实际上的标准。对于Windows，则有多个专用的管理协议，如虚拟网络计算、微软终端服务器或来自WinFrame的协议。

远程控制有其特有的优点，但是也有它的缺点。对用户来说，使用比家庭计算机更强大的公司计算机要更方便一些。而且，获得终端访问后，用户可以在远程计算机上运行任何程序，而不仅仅是WWW或FTP服务。另一个优点在于用户获得公司内部网络的用户的全部权力；而在远程节点模式中，他的权力通常都由网络管理员给以严格限制。

远程控制节约使用网络带宽，尤其是用命令行仿真时。在这种情况下，只有按键的扫描码和屏幕字符在网络上传送而不是文件或网页。

远程控制的缺点是对公司范围网络未授权访问的潜在危险。此外，对管理员来说，很难控制被远程控制的计算机资源的使用。

23.3 拨号模拟访问

拨号访问的主要思想是，使用普遍存在的公共交换电话网络（PSTN）来组织家庭计算机和安装于电话网络与计算机网络边界的RAS之间的交换连接。家庭计算机使用一个拨号调制解调器连接到电话网络，该调制解调器支持标准拨号程序并模拟电话建立与RAS的连接操作。拨号访问可以是模拟的也可以是数字的，这取决于网络提供的本地回路类型。在这一节里，我们将描述通过模拟本地回路的访问；下一节再涉及数字本地回路。

23.3.1 电话网运行的原理

最初的电话网络完全是模拟的，因为在这些网络里，终端设备（电话设备）把声波，模拟信号，转化为电路振荡，它们也是模拟信号。电话交换机也以模拟形式传送用户信息，传送方法使用第9章提到的FDM方法，把这些信号移送到另一个频带范围。

在如今的电话网络中，常使用时分多路复用（TDM）以数字形式通过PDH/SDH链路在交换机之间传送话音。然而，还是有很多模拟本地回路，它们允许相对简单和便宜的电话设备的继续使用。

图23-4展示了一个典型的电话网络的结构。这个网络由一定数量的交换机组成，这些交换机通过数字链路，很少数通过模拟链路相互连接。一般而言，网络拓扑是任意的，虽然多个低级交换机连接到一个高层交换机的多层结构也是很常见的。

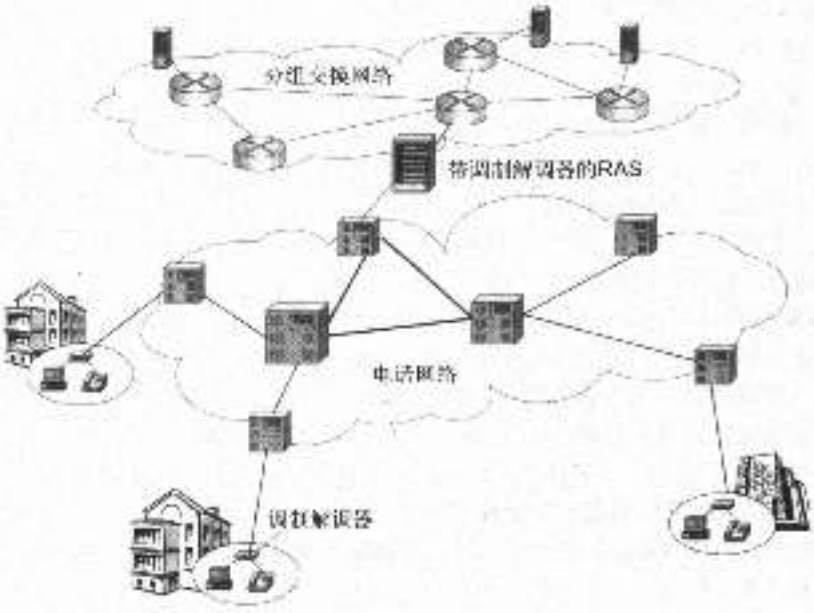


图23-4　经由具有模拟本地回路的电话网络的访问

网络用户的电话设备通过双绞线连接到接级交换机。通常，一个本地回路的长度不会超过1或2km。然而，有时通信运营商必须使用长达5～6km的本地回路。当有多个偏远用户，而为他们建立一个单独的POP又不合算时，就会发生这种情况。

电话网络，像其他电路交换网络一样，需要一个强制的连接建立程序。如果这个程序成功执行，则两个用户的终端设备之间就会建立一条可以通信的信道。这个程序被叫做信令协议。我们曾在第21章提到过这个术语，因为基于电路技术的WAN有很多地方是借鉴电话网络的。回想在模拟电话网络中，每个连接都被分配了一个4KHz的带宽。在这个带宽中，3.1KHz是用于话音传输，剩下的900Hz用于在模拟交换机之间传送信令信息并作为分配给单独用户信道之间的保护频带。

在电话网络的长期演变过程中，开发出了很多信令协议。这些协议被分为两类：用户——网络接口（UNI），运行于用户电话设备与第一个网络交换机之间的网络部分；网络——网络接口，运行于网络交换机之间。因为调制解调器是作为用户终端设备连接到电话网络的，所以它必须仅支持UNI协议。你知道这些术语是X.25网络的术语，但最初这些协议的划分是在更早的电话网络中引进的。

一台**模拟电话设备**（analog phone set）是一个简单的设备，所以，它支持的信令协议必须要简单。用户的呼叫程序通常仅包含由本地环路的电线组成的电子电路的有序的结束和中断。为了响应电路的第一次结束，电话交换机提供一定的电压给用户电路，它在电话听筒里再生为拨号音。用户在呼叫程序中扮演一个积极的角色，按需要的数字响应拨号音。

有两种发送数字到网络的方法。当使用**脉冲机制**（pulse method）时，每个数字由一定数量的频率为10Hz或20Hz的有序的开—关脉冲表示。

当使用**音频拨号**（tone dialing）——双音多频（DTMF）时，用两种音频的组合来对数字和字符

编码，一个音频来自低频组（697 770 852和941Hz），一个来自高频组（1209 1336 1447及1633Hz）。

考虑到有16种频率组合，因此不仅提供了输入数字的可能性还提供了输入如*和#这样的控制字符的可能性，如表23-1所示。

频率1 633Hz是DTMF标准的一个扩充，考虑到对A、B、C和D四个额外字符进行编码——标准的电话键盘上没有这几个字符，但是它们却被调制解调器和一些程序所使用。

音频拨号频率为10Hz，信号每50ms暂停一次。

表23-1 使用音频拨号时对数字和字符进行编码

1 209Hz	1 336Hz	1 477Hz	1 633Hz	
1	2	3	A	697Hz
4	5	6	B	770Hz
7	8	9	C	852Hz
*	0	#	D	941Hz

当使用脉冲拨号时，多个脉冲传递一个单一的数字。与此相比，音频拨号仅需一个信号来实现同一目的。所以音频拨号的速率比脉冲拨号高好几倍。

从用户电话设备收到这样一个“消息”后，电话交换机进一步转发它。如果终端设备所连的第一台交换机是一个数字交换机，它就把从用户那接收到的模拟信号转换为数字形式。

为了保证高级的处理逻辑，现代电话交换机使用**信令系统7**（signaling system 7，SS7）协议。这些协议使用分组交换技术并根据OSI模型构建，涉及从物理层到应有层的所有层。SS7协议的细节描述超过了本书的话题范围。如有兴趣，可以在电话学的教科书找到这些协议的描述。

应该注意的是电话网络仍然使用电路交换技术传送用户数据。分组交换技术仅被信令协议用于建立连接。除了SS7协议外，电话网络还使用大量的早期信令协议，包括一些模拟的协议。

23.3.2 通过电话网远程访问

为了通过电话网访问因特网或公司网络，用户的调制解调器必须拨分配给RAS调制解调器的号码中的某个。连接建立后，在电话网的调制解调器之间创建了一个带宽为4 KHz的信道。调制解调器可用带宽的确切值取决于用户调制解调器与RAS调制解调器路径间的电话交换机类型，取决于支持的信令协议类型。对调制解调器交换速率的一个主要限制是，带宽不能超过4KHz。

倘若在整个路径上至少要执行一次模拟-数字转换，则现在的调制解调器在音频拨号信道上的最佳速率为33.6Kb/s；倘若只对信息进行数字-模拟转换，那么最佳速率为56Kb/s。之所以会有这种不对称发生，是因为数字-模拟转化与模拟-数字转化相比，前者引入了相当多的偏离到传送的离散数据中。

显然，对于广泛使用的GUI和多种多媒体数据表示形式的大多数现代程序来说，这种速率难以令人满意。

RAS调制解调器通常安装于提供商的POP中。它未必是服务于特定远程用户的提供商。在因特网不像今天这样流行的20世纪80年代和20世纪90年代早期，很多大公司要保证自己雇员的远程访问。在这种情况下，RAS安装于离公司总部LAN最近的POP中，甚至就安装在公司总部里。在家办公或出差的雇员可以把他们的调制解调器连接到本地提供商的POP上，然后拨公司某RAS的调制解调器号码。对于在国外的雇员，这些拨号可能就是国际呼叫。计算机通信量通过的路径大部分是电话网络，这些访问的代价取决于距离，而距离又是电话网络常用的收费标准。

如今，因特网允许更经济地使用电话网络。连接属于某个ISP而不是某个公司的RAS时需要电话网。如果用户需要访问某个公司网络而不是因特网，则因特网用作到所求的公司网络的传送网络，该公司网络也有一个到因特网的连接。因为因特网访问的费用不取决于目的主机和到它的距离，所以对公司资源的远程访问要便宜很多，包括本地呼叫和因特网访问的费用。这种两级访问方法，虽然需要用户鉴别两次——一次是访问ISP的RAS时，一次是访问公司服务器时，但是有些

已有的协议不需要两次鉴别。一个例子是点到点隧道协议（PPTP），在这里ISP的RAS传送用户鉴别请求给公司服务器，然后在得到肯定响应后，便使用因特网连接用户到请求的公司网络。

RAS既可以使用模拟本地环路也可以使用数字本地环路连接到电话交换机。强大的RAS设备装备有几十个调制解调器，并且通常都是使用数字本地环路通过T1/E1物理链路连接的。在这种情况下，从网络到用户的数据传输不需要模拟——数字转换。从而，数据下行速率可达56Kb/s。然而，这个速率只有当路径上所有电话交换机都是数字的时候才能到达。如果至少有一台交换机是模拟的，则最大的下行速率将像上行速率一样被限制为33.6Kb/s。

23.3.3 调制解调器

虽然拨号调制解调器提供物理层服务给计算机，但它们却是实现OSI模型物理层和数据链路层这两层功能的设备。调制解调器需要一个数据链路层来检测并校正因为在数据传输期间通过电话网的位错误而引起的错误。在这种情况下，位错误的概率相对较高。所以，这个错误校正功能对调制解调器很重要。对运行于远程计算机和RAS之间的调制解调器连接之上的协议来说，数据链路调制解调器协议是透明的。它的操作只有一个效果，即，把错误率降到一个可以接受的程度。因为PPP是用于连接远程计算机与RAS的主要数据链路层协议，并且这个协议自身不能恢复丢失或损坏的帧，所以调制解调器的错误校正能力是极为重要的。

调制解调器的协议和标准定义在V ITU-T系列建议中。它们被分为下列三组：

- 定义数据传输速率和数据编码方法的标准
- 错误校正标准
- 数据压缩标准

1. 数据编码方法和数据传输速率的标准

调制解调器（modern）是最早的数据传输设备。在它们能以56Kb/s的速率运行前，它们经历了一个长期的演变。最初的调制解调器运行速率为300b/s且不能校正错误。这些调制解调器以异步模式运行，这意味着计算机传送信息的每个字节都相对于其他字节异步传送。出于这个目的，它具有与数据符号不同的起止符号。异步模式简化了设备，提高了数据传输的可靠性。然而，它也牺牲了很多信息速率，因为每个字节都补充了一个或两个冗余的起止符号。

现代的调制解调器既可以以异步模式也可以以同步模式运行。

调制解调器历史上的一个至关重要的事件是**V.34标准**（V.34 standard）的采用，它把数据传输的最大速率从它的前身V.32标准的14Kb/s提高到28Kb/s。V.34标准的一个特点是在信息交换期间对链路特征的*动态适配程序*的出现。V.34定义了10个程序，调制解调器在测试线路后，根据它们选择它的主要参数，包括载波和带宽、传输过滤器和最佳传输级别。适配发生在通信会话期间且不会终止已经建立的连接。这种适配行为的能力是微处理器和集成电路技术进步的结果。

调制解调器之间的初始连接是根据V.21标准建立的，连接的最低速率为300b/s，这考虑了对最糟糕的通信线路的操作。调制解调器持续这种协商进程直到实现了给定条件下的最高性能。与以前的标准V.32 bis相比，适配程序的使用直接使传输速率提高了两倍多。

依据线路参数的适应性调整的原理在**V.34+标准**（V.34+ standard）中得到了进一步发展。随着数据编码方法的改进，V.34+标准也提高了数据传输速率。与V.34协议传送一个代码符号平均携带8.4位相比，V.34+标准为9.8位。在代码符号的最大传输速率为3 429波特的情况下（因为音频信道带宽的限制，所以它不可能突破这个极限），改进的编码方法提供的数据传输速率为33.6Kb/s（$3\,429 \times 9.8 = 33\,604$）。

V.34和V.34+协议允许一个双线专用线路以双工模式运行。V.34和V.34+标准中的双工传输模式由两个方向的同时数据传输而不是频分复用来保证。接收到的信号等于总信号减去使用USP在线

路上传送的信号。对这个程序，要使用回波抑制程序，因为从信道的近或远端反射回来的传递信号引入了噪音到总信号中。

说明 定义了基于5种双绞线的千兆以太网技术操作的802.3ab标准所描述的数据传输方法，从V.34～V.34+标准那里继承了很多先进的特征。

V.90标准（V.90 standard）描述了这样的技术，它的目标在于给用户提供对提供商网络经济的快速的访问。这个标准保证非对称的数据交换：最大下行速率56Kb/s和最大上行速率33.6Kb/s。这个标准与V.34+标准兼容。这个标准正是我们前面在描述假设整个路径没有模拟-数字转换，但要保证下行速率为56Kb/s的可能性时，所说的那个标准。

V.92标准（V.92 standard）使调制解调器可以在连接期间接受另一个呼叫。在这种情况下，现代的电话站传送特殊的双音频信号给电话机，以便用户可以识别这种情况，并转到第二个连接（通过按电话上的Flash键），使第一个连接进入等待模式。在这种情况下，早期标准的调制解调器会终止一个连接，而这对用户来说通常很不方便；她可能正执行一个重要的任务，如从因特网下载信息（且所有他或她的工作可能会全部丢失）。

图23-5所示为一个使用模拟本地环路通过一个路由器的两台计算机或LAN之间的连接的典型结构。

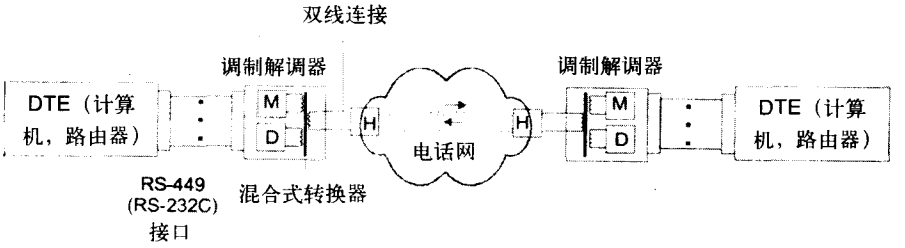


图23-5 使用拨号调制解调器连接计算机

2. 差错校正

为了通过一个异步接口与DTE一起运作的调制解调器，CCITT（现在是ITU-T）开发了**V.42错误校正协议**（V.42 error correction protocol）。在采用这个协议前，调制解调器依据Microcom公司开发的协议使用一个异步接口进行错误校正操作。这家公司已在它的调制解调器里实现了多个错误校正程序。这些程序就是著名的类2-4的 Microcom网络协议（MNP）。

在V.42标准中，另一个主要协议为：调制解调器链路访问协议（LAP-M）。然而，V.42标准还支持MNP2-4程序；所以，与V.42建议相对应的调制解调器既可以与任何MNP兼容的调制解调器建立一个可靠的完全无错连接，也可以与支持这个协议的任意其他调制解调器建立可靠的完全无错连接。LAP-M属于第22章所描述的HDLC族。一般而言，它与该族的其他协议——面向连接的、支持滑动窗口算法、帧可恢复的、带有帧编号的数据帧协议运行方法相同。它与同族其他协议的主要区别在于更多的高级协商程序，这也是LAP-M协议为几种补充的帧类型——交换标识（XID）和BREAK帧所提供的。

当建立连接时使用XID帧，调制解调器可以协商某些协议参数，如数据域的最大容量、确认的超时数值以及窗口大小。这个程序类似于PPP的协商程序。BREAK指令通知对方调制解调器关于数据流的一个临时挂断。这种情况可能会出现在与DTE的异步接口上。BREAK指令以无序号帧形式发送，它不影响数据交换会话期间的帧编号。在恢复数据传递后，调制解调器恢复数据发送，就像操作过程中没有发生过中断一样。

3. 数据压缩

实际上当通过异步接口操作时，所有的现代调制解调器都支持CCITT V.42bis和MNP-5**数据压缩标准**（MNP-5 data compression standard）（压缩比通常为1：4，虽然有些模型支持的压缩比高达1：8）。数据压缩增加了链路带宽。发送端调制解调器自动压缩数据，接收端调制解调器自动解压缩收到数据。支持压缩协议的调制解调器通常会尝试建立一个带数据压缩的连接。然而，如果另一个调制解调器不支持这个协议，那么支持数据压缩的调制解调器就转换到没有压缩的普通通信。

当调制解调器通过同步接口运行时，它们用的最多的是Motorola开发的**同步数据压缩协议**（Synchronous Data Compression，SDC）。

23.4　用ISDN拨号访问

23.4.1　ISDN的目的和结构

开发ISDN技术的主要目的是创建一个用来取代传统电话网的世界范围的网络。为了像电话网一样可用和普遍，ISDN曾被期望能提供它的数百万用户多种服务，包括电话和数据传输。因为最初没有计划通过ISDN传输电视节目，所以，开发者决定把本地环路的带宽限制为128Kb/s。

如果ISDN开发者的目的完全实现的话，那么组织家庭用户对因特网和公司网络进行访问的问题也已完全解决。然而，由于很多原因，进展过于缓慢。这个进程起始于20世纪80年代，已持续了数十年了，所以当第一批家庭用户出现时，大多数ISDN服务已经过时了。因而，如今128Kb/s的访问速率已不是一个适合所有用户的优秀的解决方案。另一个接口确保访问速率可达2Mb/s，然而，对于大多数个人用户来说，它太贵了；通常只有公司使用它来连接它们的LAN。

虽然ISDN没有成为它最初想扮演的新公共网络的角色，但是它的服务是可获得的，是想用就可以使用的。随后，我们将描述这个网络的结构和它用于组织远程访问的能力。

ISDN体系结构提供了多种服务（图23-6）：

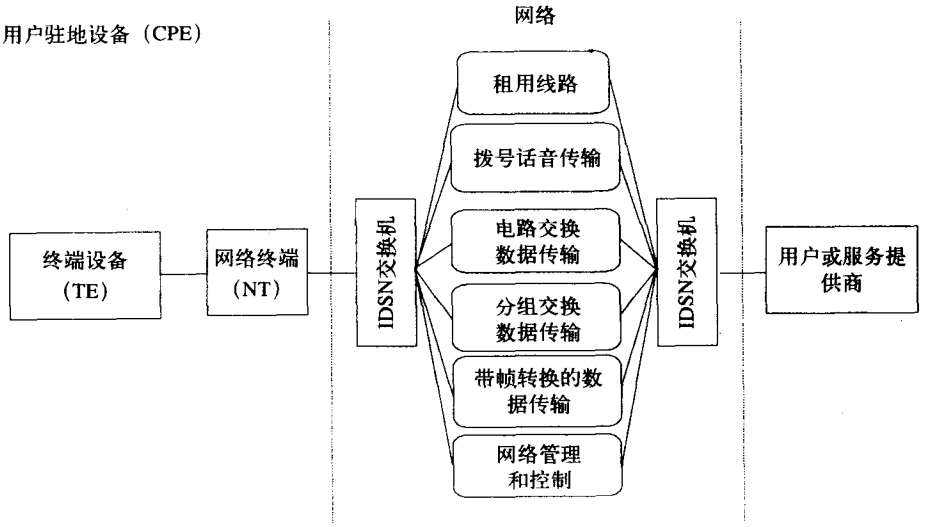


图23-6　ISDN服务

- 租用数据链路
- 通用拨号电话网
- 电路交换数据传输网
- 分组交换数据传输网

- 带帧转换的数据传输网（帧中继模式）
- 网络管理工具

由列表可以清楚地看出，ISDN传输服务涵盖了一定的服务范围，包括流行的帧中继服务。ISDN标准还描述了一定范围的应用层服务：64Kb/s的传真连接，9 600 b/s的电报通信，9 600 b/s的可视图文和多种其他服务。

所有的服务都是基于数字形式的信息传输。用户接口也是数字的；所以，包括电话机、计算机和传真机在内的所有终端设备，都必须传送数字数据给网络。DSL的组织成为阻碍ISDN普及的最严重的瓶颈，因为它要求数百万的本地环路现代化。

然而，实际上并不是每个ISDN都支持所有的标准服务。例如，虽然最初设计的ISDN结构中包含帧中继服务，但是它通常都是由帧交换机形成的独立网络实现的，并且不与ISDN交换机网络相交。

ISDN的基本速率是DS-0信道的速率（也就是64 Kb/s）。这个速率是面向最简单的话音编码方法PCM的，虽然差分编码允许有相同的质量话音以32 Kb/s或16Kb/s的速率传送。

作为ISDN基础的一个初始想法是电路交换和分组交换原理的结合使用。然而，作为ISDN一部分的分组交换网络执行一些辅助功能。这个网络用于传送信令协议的消息。因为涉及主要信息，如话音，所以它仍然通过电路交换网络传递。这种功能分布有它的明显的逻辑，因为呼叫建立连接产生了突发通信量，所以它的传输更适合分组交换网络而不是电路交换网络。

23.4.2 BRI和PRI接口

ISDN的一个主要原理是提供客户可以用来向网络请求多种服务的标准接口。这个接口建立于两种**用户驻地设备**（customer premises equipment）之间：

- **用户终端设备**（Terminal Equipment，TE），一台具有对应的适配器、路由器或电话机的计算机
- **网络终端**（Network Termination，NT），它表示一台终止与最近ISDN交换机通信链路的设备

用户接口基于三种链路类型——B、D和H。

B链路保证速率低于64Kb/s的用户数据（数字化话音、计算机数据或两者都有）的传输。数据用TDM技术分离。在这种情况下，必须使用用户设备将B信道分为多个子信道，因为ISDN通常只交换整个B信道。B信道可以使用电路交换技术来组织相当于普通电话网的租用线路的所谓的半永久连接，也可以用来连接各个用户。B信道还可以用于连接用户和X.25网络的交换机。

D信道（D channel）是以16或64Kb/s的速率传送信令信息的辅助分组交换网络的访问链路。用作到网络交换机中B信道交换机基础的地址信息的传输是D信道的主要功能。这种信道的另一功能是为用户数据的传输提供低速分组交换网络的服务。通常，当D信道不执行它们的主要功能时，网络才提供这项服务。

H信道（H channels）给用户提供384Kb/s（H0）、1 536Kb/s（H11）或1 920Kb/s（H12）的高速数据传输。在这些信道的基础上可以提供传真、视频和高质语音的高速传输的服务。

ISDN用户接口是某类预定义了传输速率的信道的集合。

ISDN支持两种用户接口——**基本速率接口**（Basic Rate Interface，BRI）和**主要速率接口**（Primary Rate Interface，PRI）。

BRI为用户提供了两个用于数据传输的64Kb/s信道（B 信道），一个用于信令信息传输的16Kb/s信道（D信道）。所有这些信道都以全双工模式运行。所以，在每个方向上BRI的总速率为144Kb/s；加上信令信息的量，则为192Kb/s。用户接口的不同信道使用TDM技术共享同一条物理双线电缆，这也意味着这些都是逻辑信道而不是物理信道。数据以帧为单位通过BRI传送，每帧包含48位。每帧包括D信道的4位，每个B信道的2字节。帧传输持续250ms，这保证了B信道64Kb/s的

数据速率和D信道16Kb/s的速率。除了数据位，帧还包含用于帧同步和保证电路信号DC分量的零值的辅助位。

BRI不仅可以支持2B+D设计，也可以支持B+D设计，或简单的D设计。

基本速率接口在I.430建议中被标准化。

PRI是提供给那些对网络带宽有很高要求的用户的。PRI要么支持30B+D设计，要么支持23B+D设计。在这两种设计中，D信道保证64Kb/s的速率。第一个变体是供欧洲使用的，第二个是供北美和日本使用的。因为2 048Mb/s数字信道在欧洲普及，在其他地区采用1 544Mb/s信道，所以把PRI多种版本标准化为一共同变体是不可能的。

也可以组织B数目更小的PRI变体；例如，20B+D。B类型信道可以组合为一个保证总速率最高可达1 920Kb/s的高速逻辑信道。当多个PRI安装在同一个用户的驻地时，它们可以拥有一个共同的D类型信道，并且在这个缺少D信道的接口上，B信道的数量可以增加到24或31。

PRI可以是基于H型信道的。在这种情况下，接口的总带宽仍然不能超过2 048Mb/s或1 544Mb/s。对于H0信道，它有可能使用3H0+D接口（北美变体）或5H0+D接口（欧洲变体）。对于H1信道，则有可能把接口组织为包含一个D信道和一个北美变体的H11信道（1 536Mb/s）或一个欧洲变体的H12信道（1 920Mb/s）的接口。

PRI帧为T1或E1信道准备了DS-1帧结构。

PRI在I.431建议中被标准化。

说明　B信道和D信道都是本地环路的逻辑信道。物理上，本地环路是一根单一的双绞线。D信道和B信道是这根双绞线使用TDM技术为这个物理介质组织的。

23.4.3　ISDN协议栈

在ISDN中，有两个协议栈：D信道栈和B信道栈（图23-7）。

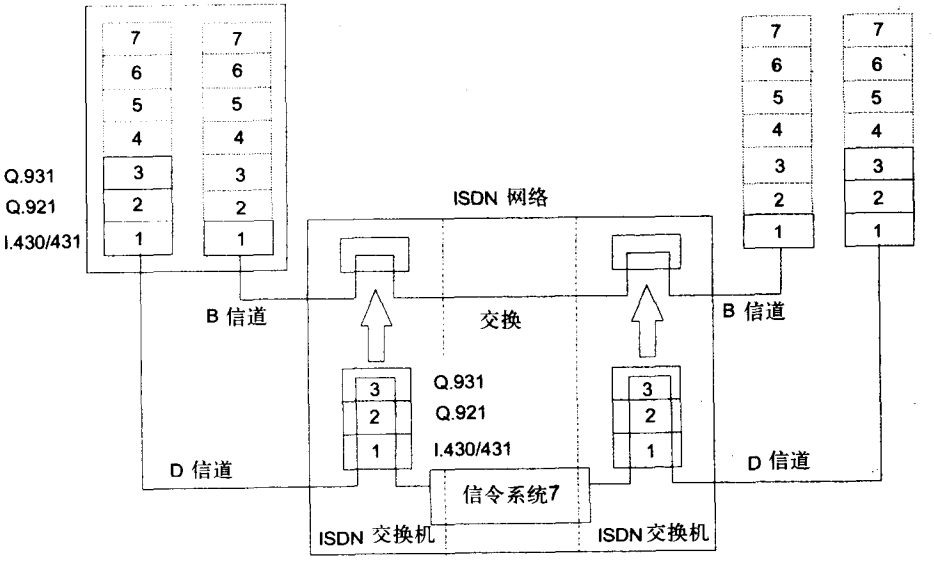


图23-7　ISDN结构

在ISDN内由*D信道形成的网络*用作信令消息传输的传输分组交换网络。X.25网络技术用作该网络的一个模型。对由D信道形成的网络，定义了下列三个协议层：

- 物理层协议由I.430/431标准定义。

• D信道上的LAP数据链路层协议由Q.921标准定义。

• 网络层可能会使用Q.931协议（它执行电路交换服务的用户的呼叫任务）。

B信道形成一个带数字电路交换的网络，它传送用户数据——即，数字化的话音。根据OSI模型，在ISDN交换机的B信道上只定义了物理层协议，即I.430/431协议。对B信道而言，电路根据从D信道收到的指令进行交换。当Q.931协议的帧由某交换机选择路由后，呼叫者到被叫用户的电路的下一部分也同时被交换。

LAP-D属于已提过多次的HDLC族，LAP-D拥有该族的所有通用特性，它也有一些自己的特征。LAP-D帧的地址由两个字节组成，一个字节定义封装入帧的分组所要发送给的服务编码，另一个字节对某个终端地址（如果有多个终端连接到同一本地环路）。终端设备可以支持多种服务，包括依据Q.931协议的连接建立、X.25分组交换和网络监控。LAP-D保证了两种操作模式：面向连接的和无连接的。后一种模式可用于网络监控和管理。

Q.931协议是ISDN在UNI部分使用的信令协议。该协议在其分组中带有被呼叫用户的ISDN地址，基于该地址以调整交换机来支持适当的B-类电路。图23-8说明了按照Q.931协议建立连接的过程。

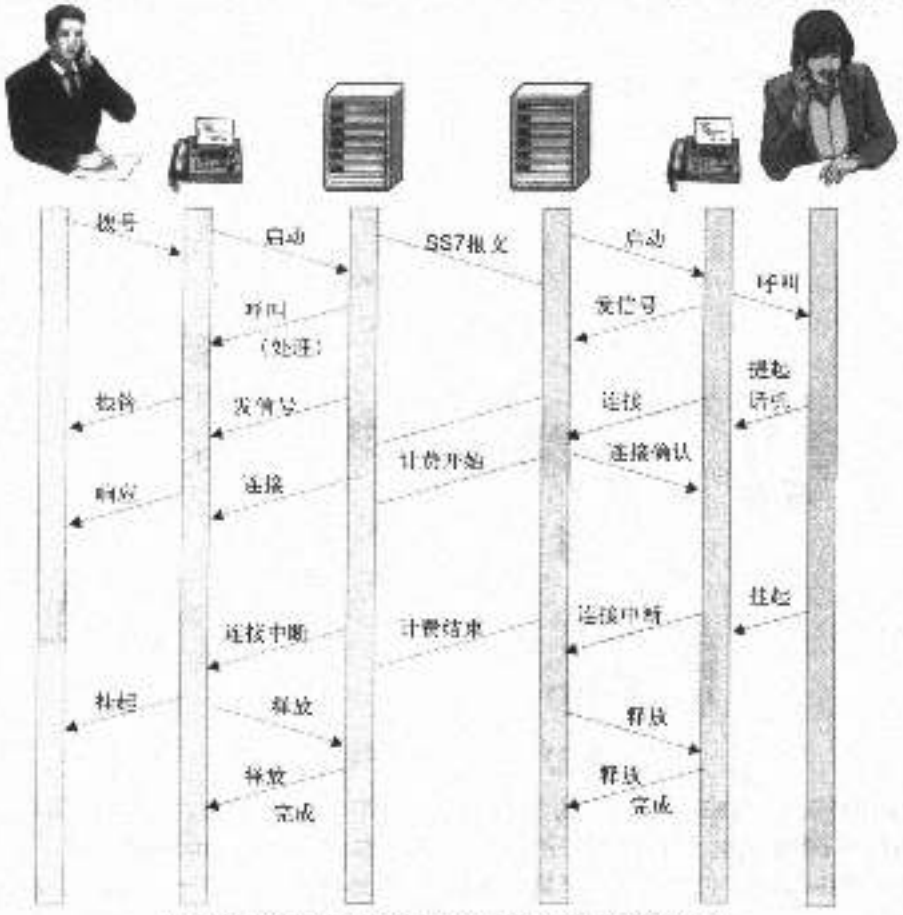


图23-8 在ISDN中依据Q.931协议的基本连接建立程序

在用户拿起话筒并拨被叫用户号码后，ISDN电话形成建立分组并通过D信道把它发送给连接的ISDN交换机。交换机以呼叫处理分组响应用户电话设备，当这个分组到达时，用户电话就产生"嘟嘟"长音。同时，交换机记住连接建立请求的这个行为，并把收到的消息传递给下一台交换机，这台交换机的地址是它在一张类似于分组交换网络路由器的路由表中找到的。同时，Q.931协议的报文

被转换为SS7协议的初始地址报文（IAM）（图23-8没有具体描述SS7报文）。当SS7报文通过网络传播时，它们在传送交换机中设置连接就绪状态。被叫用户设备所连的网络输出交换机，把SS7协议的IAM报文转换为Q.931协议的呼叫报文。收到这个报文后，被叫用户的电话开始响铃。如果用户拿起听筒，那么电话就会产生连接报文，它经过所有的传送交换机沿着相反的方向传播（自然以SS7报文格式）。当连接报文穿越这个网络时，所有的传送交换机通过合适的方法交换B信道建立已连接状态。

任何ISDN终端设备都必须支持Q.931，所以，ISDN电话设备比模拟电话设备要复杂的多。由图23-8可以清楚看出，ISDN把Q.931报文转换为SS7报文，然后在本地环路中执行一个相反程序。

23.4.4　用ISDN进行数据传输

尽管与模拟电话间有很多不同，但ISDN使用的方法却与之类似——更确切地说，就是更快的模拟网络。例如，BRI建立了一个速率为128Kb/s的双工数据交换模式（两个B信道的逻辑组合）。PRI则允许2.048Mb/s。此外，数字信道的质量也比模拟信道的高很多。这意味着损坏帧的比例将会相当的低，同时数据交换的有效率也将极大提高。

通常，BRI用于连接个人计算机或小型家庭LAN的通信设备。PRI则用于通过某路由器连接中型LAN。

通过ISDN的远程访问方法，如图23-9所示。

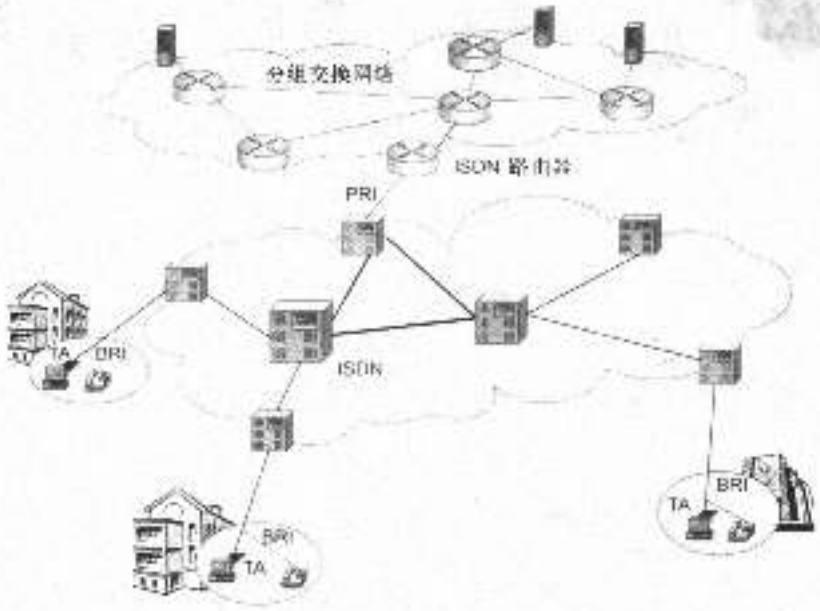


图23-9　使用ISDN进行远程访问

图23-10展示了依据ITU-T开发的方法连接到一个ISDN的用户设备。所有的设备都被划分为几个功能组，从于特定的组、右多个用于组间设备互联的**参考点**（reference point）。

端用户的终端设备（TE1）形成了第一个设备功能组，它可能是一台数字电话或一台传真机。S参考点对应于一台独立的终端设备——网络终端（NT1功能组）或用户接口的集中器（NT2功能组）的连接点。根据规定，TE1支持一个SDN用户接口——BRI或PRI。

如果用户的TE1是通过BRI连接的，那么数字本地环路是依据双线设计实现的，类似于一个普通的模拟电话网的本地环路。在这种情况下，在ISDN网络连接点（参考点U）的DSL部分的数字编码可能采用2B1Q编码。DSL双模式由通过同一条双绞线的两个方向的同时的信号传输组织，且

响应取消控制并从总信号中减去传送的信号。这种情况下本地环路的最大长度为5.5km。

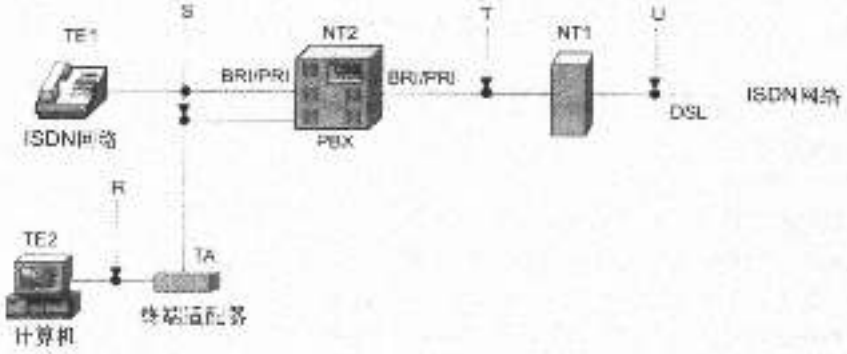


图23-10 连接用户设备至ISDN

使用PRI时，DSL代表一个T1或E1信道，这意味着它表示一个总长约为1 800m的四线连接。相应地，对DSL，使用其他PRI编码，HDB3（在欧洲）或B8ZS（在北美）。

与Tb1相比，TE2功能组的设备不支持BRI和PRI。TE2功能组的设备可能是计算机或有连续接口的路由器而不是ISDN，如RS-232C、X.21或V.35，为了连接这些设备到ISDN上，有必要使用一个**终端适配器**（terminal adapter，TA）。计算机的TA是以网络适配器的形式制造的。参考点*R*对应于TE2设备到终端适配器（TA）的连接点。本地环路的类型不取决于终端设备是否通过TA或是否直接与网络通信。

*NT2*功能组的设备是用于集中和多路复用用户接口的数据链路层或网络层设备。这种类型的设备有：交换多个BRI的办公PBX、以分组交换模式运行的路由器（如，通过D信道）和复用多个低速信道为一个B信道的简单TDM多路复用器。NT2设备到NT1设备的连接点是参考点*T*，与NT1设备相比，这种设备的存在不是必须的，因为这个原因，参考点S和T被合并指派为参考点*S/T*。物理上，点S/T上的接口是四线的线路。对BRI，选择的编码方法是双极性交替信号反转码，0由0电压表示，1[⊖]由电压转换表示。对PRI，使用的是其他编码，明确地说，与T1和E1使用的编码相同（如，对应的B8ZS和HDB3）。

连接用户设备与ISDN网络的NT1功能组设备（Devices of the functional group）是协调BRI[⊜]或PRI与数字本地环路（DSL）接口的物理设备。事实上，NT1是一个协调编码方法、在使用中的线路编号和电路信号参数的CSU设备。参考点*U*对应于NT1设备与网络的连接点。

> **说明** NT1设备可以为通信运营商所有（虽然它通常安装于用户场地）；也可以选择为用户所有。通常，在欧洲NT1设备被认为是网络设备的一部分；所以，用户设备（如配有ISDN接口的路由器）创建时是不嵌入NT1设备的。在北美，NT1设备被认为是用户设备的一部分；所以，用户设备通常都配有嵌入式的NT1设备。

因此，为了组织远程访问，就必需为用户计算机装备TA并在POP安装一个至少装备了一个PRI的路由器。在这种情况下，用户个体的最大访问速率将等于两个B信道的速率（如128Kb/s）。ISDN TA的驱动器可以把两个独立的B信道组合为一个逻辑信道。为了这个目的，使用了一个特殊的PPP扩展，即著名的多链路PPP（RFC1990）。

如果远程用户对64Kb/s的访问速率感到满意，则该用户可以使用BRI的第二个B信道进行ISDN电话的并行操作。注意，当使用模拟拨号调制解调器时，这是不可能实现的。

⊖ 原文为0，译者认为应为1。——译者注

⊜ 原文为BPR，译者认为应为BRI。——译者注

23.5 XDSL技术

在20世纪90年代中期，出现了数字ISDN本地环路的另一选择。这一族的技术就是著名的xDSL。xDSL包括：

- 非对称数字用户线（Asymmetric Digital Subscriber Line，ADSL），它在通信运营商的广告中经常被称做宽带访问。
- 对称DSL（Symmetric DSL，SDSL）
- 速率自适应DSL（Rate Adaptive DSL，SDSL）
- 甚高速DSL（Very high-speed DSL，VDSL）

以ADSL为例，考虑xDSL的主要操作原理。这个技术是最流行的技术，因为它是为最广大的需要访问因特网或通过因特网访问他们的公司网的家庭用户开发的。

ADSL访问，类似于拨号访问，使用电话本地环路和调制解调器。然而，ADSL和拨号访问的主要不同点在于，ADSL调制解调器仅运行于本地环路上，而拨号调制解调器通过多个传送交换机建立连接来使用电话网。

与保证通过信道的数据传输的带宽为3 100Hz的传统拨号调制解调器（如，V.34和V.90）相比，ADSL调制解调器带宽约为1MHz。受ADSL调制解调器支配的实际带宽取决于用户驻地到POP的电缆长度及使用的线路的横截面。

ADSL访问方法如图23-11所示。除了ADSL访问忽略了电话机以保证对计算机和电话的访问外，这个方法与图23-2所示的使用通用本地环路的一般方法很接近。

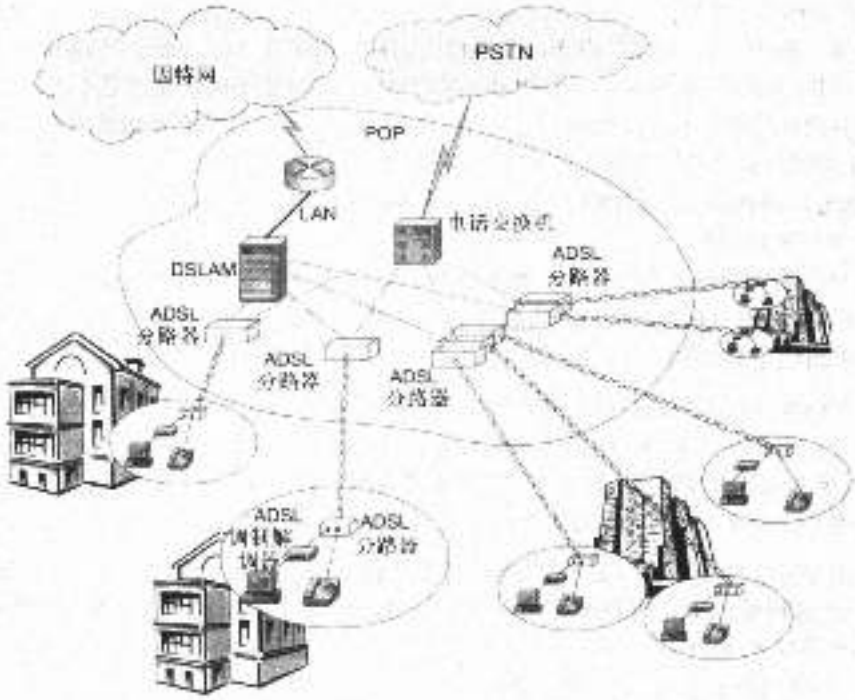


图23-11　ADSL调制解调器与传统调制解调器的差异

连接用户和POP的某条电话线路的两个终端的ADSL调制解调器构成了三个信道：用于网络到计算机数据传输的高速下行信道，用于计算机到网络的低速上行信道和用于传统电话通信的电话信道。在网络-用户信道中的数据传输速率为1.5Mb/s到6Mb/s，用户-网络信道传送数据速率为16Kb/s

到1Mb/s，电话信道带宽仍为传统的4KHz（图23-12）。

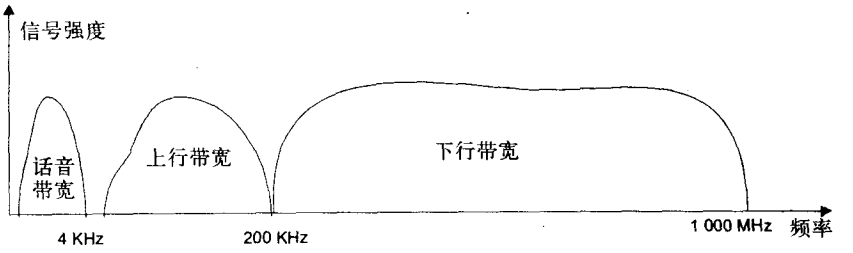


图23-12 ADSL信道间的本地环路带宽分布

为了保证下行通信量和上行通信量间的非对称速率分布，信道间的带宽是非对称地划分的。图23-12展示了信道间带宽的近似分布。之所以说信道速率是近似的是因为带宽的确切值事先是不知道的。它取决于本地环路的长度、所用线路的横截面和双绞线的质量。此外，信道间的带宽分布不仅依赖于技术上的可行性和调制解调器的功能性，还依赖于提供商的意愿。通常，ADSL调制解调器允许可以改变带宽分布和数据各方向数据传输速率的调节。

在用户驻地，装有一个分路器，它分配ADSL调制解调器和普通模拟电话的频率，从而保证它们的共存。

在POP中，必须安装一个名为**DSL访问复用器**（DSL access multiplexer，DSLAM）的设备。它接收分路器从本地环路另一终端的话音数据中分离出来的计算机数据。在DSLAM中的ADSL调制解调器的编号必须对应于使用提供商电话本地环路的远程用户的编号。

在把信号调制为离散形式后，DSLAM把数据发送给通常位于POP驻地的IP路由器。此外，数据还被提供给提供商主干，并按照它们的目的IP地址递送出去——要么递送给公共因特网网址要么递送给用户私有网络。被分路器分离的话音信号被传递给某个电话交换机，由它处理这些信号就好像用户本地环路直接连接到这个交换机一样。

xDSL技术的广泛使用必须伴随有ISP和电话网提供商中的某一个操作重组。这是因为它们的设备必须协同运行。另一可能的变体是，某个有竞争力的通信运营商从传统通信运营商处租用了大量的本地环路或租用一些DSLAM调制解调器。

G.992.1标准描述了ADSL调制解调器收发器的操作。ADSL支持多个信息编码的变体：DMT、CAP和2B1Q。xDSL技术的实现在许多方面依赖于编码技术的实现，这是因为DSP的使用增加了数据传输速率，同时还增加了调制解调器与DSLAM间的距离。

ADSL信道的速率依赖于物理线路的质量以及调制解调器与DSLAM的距离。距离越大，访问速率越低。通常调制解调器降低了传输速率；所以，当在一个具体本地环路安装调制解调器时，会选择一个可以保证最大传输速率并满足传输质量的最佳运行模式。

ADSL调制解调器的高速率给服务提供商带来了一个新问题，即带宽短缺。如果每个ADSL用户都以最高速率——比如，1Mb/s从因特网上下载数据——那么服务100个用户的提供商就需要一个100Mb/s的信道（高速以太网）。如果用户允许以6Mb/s的速率运行，则提供商将需要一个622Mb/s的ATM或千兆因特网信道。为了保证请求的速率，很多DSLAM都有一个嵌入式ATM或千兆以太网交换机。ATM技术之所以吸引DSLAM开发者，是因为它的高速率及面向连接的特性。在数据链路层使用ATM时，在开始数据传输前，用户计算机要建立一个到提供商网络的连接。这保证了控制用户访问及根据连接时间和传递的数据向用户收费的可能性——如果SLA考虑了这些参数的话。

*SDSL技术*允许在同一对用户本地环路上组织两个对称的数据传输信道。通常，在这种情况下不提供信道频率调节，上行和下行信道速率都为2Mb/s。然而，与ADSL技术类似，这个速率依赖于线路质量和到DSLAM的距离。SDSL是为LAN包含如网址或数据库服务器这样的内部信息资源

的小型办公室设计的。所以，这种情况下就要求用对称通信量，因为SDSL访问不仅可以用于访问外部网络也可以用于保证从外部到这些服务器的访问。

xDSL访问的广泛使用给ISDN制造了很大的麻烦。当使用这种类型的本地环路时，主要是通过两种网络的结合：计算机和电话。用户也可以获得综合服务。然而，这两种网络的存在对于用户而言是透明的。因为他们察觉到的唯一的一件事是可以同时使用标准电话和连接到因特网的计算机。因为计算机访问速率的关系，这可能超过了PRI ISDN的能力。同时，因为IP网络结构的低费用，它的费用也相当的低。

23.6　用有线电视访问

有线电视是一个有专门的分布式的用户本地环路基础设施的电信服务。虽然CATV网络没有电话那么普遍，但是在一些工业国家连接用户和提供商POP的同轴本地环路数量也可以与电话本地环路相媲美了。我们知道同轴电缆有相当大的带宽（至少700MHz），所以CATV本地环路可以同时处理电话、计算机和TV通信带。

你已经了解到把CATV本地环路作为访问因特网、电话网和CATV网络的通用本地环路的通常方法。图23-2是CATV本地环路的一个用例。现在，详细研究这种访问。

CATV本地环路的不同在于多个用户通过wired OR设计同时连接到同轴电缆（图23-13）。它们可能是多个总宅或公寓建筑里的几十个甚或几百个公寓。所以，CATV本地环路是一个典型的共享介质，如同同轴以太网中使用的一样。

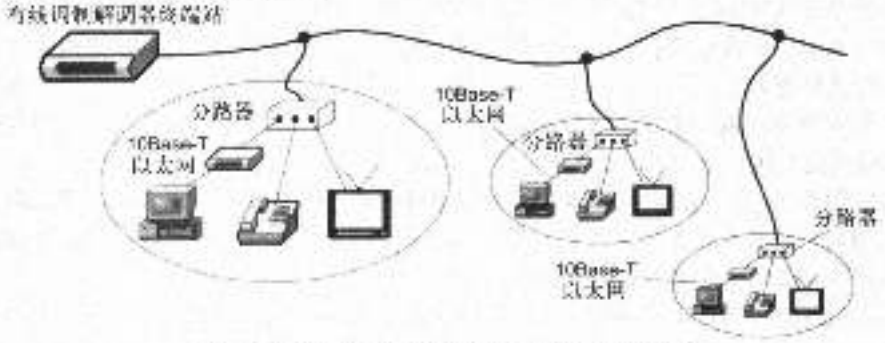


图23-13　连接有线调制解调器到CATV本地环路

如果没有与有线调制解调器连接，CATV设备就用于从位于提供商POP处的信息资源到有线TV用户的TV广播。为了这个目的，使用了一个从50MHz到868MHz的频率范围（这个范围的具体边界取决于当地国家频率分配政策）。每个有线TV节目都从这个范围里分配一个6MHz到8MHz的带宽，这个带宽的信号是加密了的，但它可以被那些预定了这个节目的用户的TV机解密。

为了使用这些本地环路，每个用户驻地都配有一个分路器和一个有线调制解调器。在被称**有线调制解调器终端站**（cable modem termination station，CMTS）的POP站，装有主调制解调器。

对于双向数据传输，用户有线调制解调器和CMTS使用未被TV节目使用的可用频率。通常，这是一个低于TV范围的频率相对较低的从5MHz到50MHz的范围，以及一个550MHz以上的高频范围。

低频范围给速率较低的上行信道使用，高频访问供高速下行信道使用。数据传输速率在上行方向可达10Mb/s，在下行信道可达30～40Mb/s。用户调制解调器仅可以和CMTS通信。

因为下行和上行信道使用不同的频率，所以CATV本地环路形成了两个共享介质。

对下行信道，CMTS是唯一的信息传递者，所以它没有对介质访问的竞争。CMTS通过以太网寻址和时分划分，使用下行信道传递所有的用户帧。

上行信道被连接该本地环路的所有有线调制解调器用于多访问模式。在这个共享介质中，CMTS扮演着仲裁器的角色。每个用户调制解调器只有在获得CMTS准许后才能传输数据。为了防止调制解调器使用信道时间过长，CMTS给每个用户分配了一个有限时间槽。时间槽仅分配给活动的调制解调器，这使得有限的带宽得以最大程度地利用。对新连接的调制解调器，分配以特殊的时间槽。当某个用户的调制解调器获得连接后，它使用它的一个时间槽通知CMTS它在网络中的存在。然后，它将使用分配给它的时间槽。

用户有线调制解调器可以有一个连接传统电话机的连接器，它也在低频范围分配了一个4MHz的带宽。在这种情况下，用户将获得三种访问——电话、计算机和TV——从同一个提供商处。

23.7 无线访问

在前面的几章中，我们讨论了无线访问的具体特征。在第9章我们提到无线通信的一般原理，在第14章，我们描述了无线LAN和PAN技术。无线数据传输广泛用于组织远程访问，尤其是当提供商不能保证客户有线访问时。这种情况最常见于没有连接客户私宅到自己POP的本地环路的有竞争力的提供商。另一个典型情况是为了某个会议而进行的，对某具体建筑的临时高速无线访问的组织（如，在没有配有可以保证所需速率的有线访问的酒店）。

无线访问既可以是固定的也可以是移动的。

*固定无线访问*是为计算机位于一个限制区域的用户组织的，大多数情况都在同一栋建筑里。在这种情况下，提供商可以使用一个定向天线和一个预定好强度的发送器来保证在一个限制的覆盖区域内，如一栋建筑，高频信号的稳定接收。如果提供商有大量的固定无线访问用户，那么它将使用多个定向天线来覆盖用户所在的所有区域。

固定无线访问的另一个广为人知的名字是**无线本地环路**（wireless local loop，WLL）。这个术语反映出，尽管没有电缆，用户还是被限制在某一特定的地理位置。

目前有*窄带*WLL和*宽带*WLL。最初类型的WLL不保证TV信号的传输。它们仅支持相对低速的计算机访问（64～128Kb/s）和电话信号的传输。第二种类型的WLL通常是基于TV广播信号的；所以，它们运行在高频范围，并保证所有的三种访问。在这种情况下，计算机数据以每秒几十万比特或每秒几兆比特的速率传递。

最新类型的系统包括**多信道多点分布服务**（multichannel multipoint distribution service，MMDS）和**本地多点分布服务**（local multipoint distribution service，LMDS）。MMDS运行频率约为2.1GHz，LMDS在北美运行频率约为30 GHz，在欧洲约为40GHz。两个系统都保证用户TV、电话和计算机服务的双向数据传输。因为MMDS运行频率相对LMDS低，所以它保证了相对较大的覆盖区域。有向MMDS天线的一极通常可以服务半径50km的区域。LMDS发送器的半径不超过5km，在城市中则更小。另一方面，LMDS可以保证用户相当高的访问速率，可达155Mb/s。

窄带WLL和宽带WLL使用不同的信号复用方法，这既是为了保证在天线方向区域内的用户的同时操作也是为了分离TV、电话和计算机通信量。通常，这里会结合使用FDM和TDM。例如，对每种通信量，可以根据FDM原理分配一定范围的频率。然后，在为计算机通信量分配的这个频率范围中，还可以使用预定义好的访问算法的异步TDM来共享介质，比如，使用中心仲裁器。对需要有保障带宽的用户，可以使用同步TDM，由它形成无线PDH/SDH信道。

不幸的是，WLL技术在很多方面是专用的，这意味着它们与访问设备和中心站不兼容。为了消除这些缺点，开发了IEEE802.6标准，它定义了使用频带、复用方法和提供的服务的一些一般原则。这个标准允许基于FDM的多种复用方法，以及基于同步和异步TDM。这考虑了不同WLL设备制造商的利益并保证了这些系统的最大灵活性。

802.11技术也可以用于组织固定的无线访问。然而，它不常用于这个目的，因为它是专门面向

于计算机访问的，它不关心电话和TV通信量的特征，即，以一固定比特率提供访问的可能性。用在802.11里的CDMA/CA访问方法不能保证所需的实时通信量的QoS等级。

然而，一些提供商使用802.11技术为满足于无保障带宽因特网访问的用户的服务。对用户临时逗留的如机场或火车站这样的区域所使用的临时访问来说，这个技术也是很受欢迎的。

如今对因特网的无线访问由移动电话网络提供。2G移动电话像使用分组传输一样使用**通用无线分组服务**（General Packet Radio Service，GPRS）协议来提供低速因特网访问。GPRS仅运行于D-AMPS和GSM网络。这种访问的最大速率很低，只有2 400～9 800Kb/s。在最近已经开始开发的3G移动网络中，这个速率将有较大增加，会达到2Mb/s。

小结

- 当需要保证家庭用户或小公司的雇员到因特网或公司网络的访问时，就需要使用术语远程访问。
- 远程访问的客户有多种，他们的不同点在于使用的本地环路，是否可以使用家庭LAN，设备的访问速率及需要提供访问的资源类型——因特网公共域资源或公司私有网络资源。
- 提供商尽量使本地环路通用化——也就是说，主要的三种类型终端设备的通信都可以通过它的传输量这三种设备是：电话机、电视机和计算机。
- 远程访问的基础服务是远程节点方式，这个方式下用户的计算机成为提供商LAN或公司私有网络的一个节点。
- 远程控制是远程访问的一个特殊方式，在这个方式下，用户计算机仿真所连接的另一台计算机的终端。远程访问运行用户获得对另一台计算机的完全控制权，并可以在它上面运行任何程序。这对使用者很便利但却给公司资源带来了潜在危险。
- 最早类型的远程访问是通过PSTN模拟本地环路的拨号访问。使用一个拨号调制解调器，用户计算机就可以建立一个连接到分组交换网RAS的连接。
- 分配给电话网用户的固定的4KHz带宽限制了拨号调制解调器的速率。V.90调制解调器保证了可达33.6Kb/s的上行速率和可达56 Kb/s的下行速率。然而，56 Kb/s的下行速率只有当用户到RAS之间的所有传递电话交换机都是数字时才能得到保证。
- 开发ISDN技术是为了创建一个保证对传统服务进行补充的计算机网络服务的通用网络。然而，如今对于大多数要求访问多媒体信息的客户来说，它的128 Kb/s的访问速率实在是太低了。
- ADSL技术充分利用了电话本地环路带宽，它把它分为三个信道——一个二元话音信道，一个保证传输速率可达 1 Mb/s的上行信道和一个保证计算机数据传输速率可达6Mb/s的下行信道。对电话用户的4KHz的带宽限制不影响ADSL调制解调器的运行，因为计算机数据被分离出来并直接送往最近的POP处的分组交换网络。
- 有线调制解调器专用于同轴CATV本地环路，是一个为连接到同一个电缆的多个用户服务的共享介质。同轴电缆的带宽保证了可达10Mb/s的上行速率和可达40Mb/s的下行速率。
- 固定无线访问使用了大量的专用技术来传送电话、TV和计算机信息给用户。 为了提供多种服务，这些访问综合使用了FDM、TDM、分组交换和电路交换技术。
- 移动访问是通过2G移动电话网络的低速数据传输的一个流行的补充服务。3G移动网络标准允诺有更高的传输速率，但它们才刚刚开始投入使用。

复习题

1. 决定组织远程访问的复杂性的因素有哪些?
2. 没有有线本地环路的提供商如何保证用户的网络访问服务?

3. 当为客户组织远程访问时，必须要考虑它们的哪些属性？
4. 什么样的本地环路可以被认为是通用的？
5. 远程访问用户的PC在什么网络中？
6. 远程节点方式和远程控制方式的不同点是什么？
7. 当配置远程路由器时使用的是什么访问方式？
8. 为什么拨号调制解调器的速率要比ADSL和有线调制解调器的访问速率低很多？
9. 调制解调器和DSU/CSU设备的不同点有哪些？
10. 假设你确定你的调制解调器在专门的双线线路上无论是同步方式还是异步方式都是稳定的。你会选择哪种方式，为什么？
11. 可以把拨号调制解调器归为哪一层设备（根据OSI模型术语）？
12. DSLAM的功能是什么？
13. 对拨号服务提供商LAN和ADSL服务提供商LAN的要求的区别有哪些？
14. 有线调制解调器使用的访问共享介质的方法是什么？
15. 可以使用同一根同轴电缆保证同一栋公寓建筑里的住户的访问吗（多于400间公寓）？
16. 为什么802.11技术很少用于组织固定的无线访问？
17. MMDS和LMDS技术的区别在哪？
18. 为什么大多数家庭用户都对移动无线访问不满意？

练习题

1. 假设你买了一个V.90调制解调器，并尝试使用电话网建立一个与同样使用V.90调制解调器的同事连接。你确定在你和你同事之间的所有电话交换机都以数字模式运行。你将以什么速度连接？
2. 如果两台PC是使用TA连接到网络的，哪种ISDN服务可取？如果它们需要经常以2 400b/s的速率交换速率且突发速率可达9 600b/s呢？如果分组延迟不重要呢？
3. 如果两个LAN是通过路由器连接到这个网络，且互联网通信量速率在很长时间内为100Kb/s到512Kb/s，则哪种ISDN服务可取？
4. 在哪种情况下组织远程访问通过接口为B+D的ISDN更可取？在哪种情况下使用64Kb/s租用数据线路会更好？什么时候使用CIR等于64Kb/s的永久帧中继虚电路更有利？
5. 如果运行于给定本地环路的ADSL调制解调器错误率极高，则需要改变哪些东西？

第24章 安全的运输服务

24.1 引言

在本章，也是本书的最后一章里，我们将涉及一些流行的安全运输服务。这些服务允许像这样的流量运输：它以一种安全的方法使用如因特网这样的公共网，并保证所传输信息的真实性、安全性和机密性。

提供这种服务的最简单的技术是受保护的信道技术，它保证对“点到点”拓扑的公共网络用户之间流量的保护。这种保护运用了使用多种用户认证和流量加密方法的所有工具。在IP网络中，下列两项技术被广为使用：安全套接字层（SSL）和因特网协议安全（IPSec）。SSL运行于OSI模型的表示层，这使得它对应用程序是非透明的，因为它们必须为流量保护使用明确的API请求。IPSec是更通用的工具，因为它运行于网络层；因此，它对应用程序是绝对透明的。当使用IPSec时，不需要重写应用程序。

一个更强大的流量保护工具是虚拟专用网（VPN）。VPN是一种“网络中的网络”，这意味着VPN是一个为某公共网络的用户创建了一个表面上的专用网服务。VPN所模拟的专用网的最重要特征是保护VPN用户流量不受公共网络用户的攻击。除了模拟专用网的这个特征外，VPN还可以提供用户使用专用地址空间的能力（如属于10.0.0.0 IP网络的私有地址）。此外，VPN还可以提供近似于租用线路服务所提供的QoS。

VPN的基础技术可以分为两类：使用数字加密的技术和在流量分离的基础上确保安全的技术。第一类技术使用了安全信道技术，用它们来连接任意数目的用户网络，而不仅是两个。IPSec VPN是这类VPN的代表。

另一类VPN使用永久虚电路技术，它保证用户流量与其他网络用户流量的可靠的分离。基于流量分离的VPN不使用加密，因为PVC原理排除了来自其他PVC所连接用户的外部攻击。这类VPN基于ATM、帧中继和MPLS技术。而ATM VPN和帧中继VPN仅是这些技术的PVC服务的别名。这些服务曾在第21章提到过。本章主要集中于介绍这些新类型的VPN的新类型，即，MPLS VPN，它极大地拓广了VPN服务的范围。

24.2 IPSec受保护的信道服务

IPSec服务的主要目标是确保使IP网络进行安全数据传输。IPSec的应用范围包括数据完整性、真实性和机密性。实现这些目标的基础技术是加密。

对于目的在于实现这个目标的协议来说，最常用的术语就是——**安全信道**（secure channel）。术语信道强调了在连接两个网络节点（主机和网关）的路径上的数据保护是有保障的[⊖]。

24.2.1 受保护的信道的服务层次

IPSec是公共网上众多可用的安全数据传输技术中最流行的一个。安全信道可以使用实现于OSI模型各层的嵌入式OS工具来创建（图24-1）。

如果数据保护是使用OSI模型上面几层（应用层、表示层或会话层）来执行的，那么这种保护的实现方法是独立于数据运输的基础技术的（例如IP、IPX、Ethernet或ATM）。毫无疑问这个特征

⊖ 受保护的信道的特征在第6章曾简要讨论过。

是这种方法的优点。另一方面，这种情况下的应用程序依赖于具体的安全协议，因为它们必须有对安全信道协议函数的明确的内在的调用。

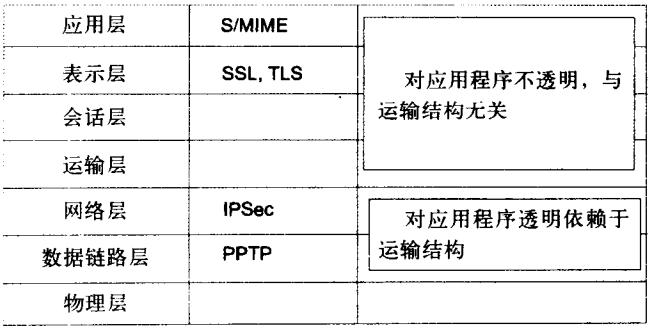


图24-1 创建安全信道所涉及的不同的OSI模型层的协议

实现于最高层应用层的安全信道仅用来保护某种特殊的网络服务，如文件、HTTP或邮件服务。因而，S/MIME协议专门用于电子邮件（e-mail）的保护。当使用这种方法时，需要为每个服务开发一个单独版本的保护协议。

流行的**安全套接字层**（Secure Socket Layer，SSL）协议和它的被称为**运输层安全**（Transport Layer Security，TLS）的开放式实现，保护任何应用层协议或单独的应用程序的PDU。显然，这些协议是比应用层安全协议更通用的保护协议，因为任何应用程序都可以使用它们。然而，为了实现这个目的，必须重写应用程序以包含对安全信道协议API函数的具体调用，这运行于表示层。

当安全信道工具保护网络层和数据链路层协议的帧时，它们就对应用程序不可见。此时，开发者必须面对另一个问题，即，安全信道服务对下层协议的依赖性。例如，点到点隧道协议（PPTP），通过把运行于数据链路层的点到点协议（PPP）的帧封装进IP分组来保护这些帧。同时要注意，PPTP本身不属于数据链路层。一方面，这使得PPTP服务通用化，因为安全信道服务的用户可以使用任何网络协议，如IP、IPX、SNA或NetBIOS。另一方面，这种方法暗含了对网络部分的数据链路层协议使用的严格要求，这个数据链路层协议用于提供用户对安全信道的访问。当使用PPTP时，*在数据链路层只可以使用PPP*。虽然PPP广泛用于链路访问，但是它也有强劲的对手，如既用于LAN又用于WAN的千兆以太网和高速以太网协议。

运行于网络层的IPSec是一个折衷的变体。它对应用程序通明，但实际上它可以运行于任何网络，因为它是基于普遍的IP的，并且它可以使用任意数据链路层技术，包括PPP、以太网和ATM。

24.2.2 IPSec协议间的功能分配

在因特网标准中，IPSec被称为*系统*。IPSec是开放标准的一个协商集。它有一个组织良好的核，这个核是很容易由新功能和新协议实现的。

下列三个协议组成了IPSec的核：

- **认证头部**（Authentication header，AH）——这个协议确保数据完整性和真实性。
- **封装有效载荷**（Encapsulation security payload，ESP）——这个协议对传送的数据进行加密，因而保证机密性。它也可以支持数据认证和完整性。
- **因特网密钥交换**（Internet Key Exchange，IKE）——这个协议解决了给安全信道的端节点提供所需的用于数据认证和加密协议的私钥这个附属问题。

正如从协议功能简介里看到的，AH和ESP的功能部分重叠。与仅负责保证数据完整性和真实性的AH相比，ESP可以给数据加密并执行一些AH协议功能。然而，后面将会提到，它的与保证数据认证和完整性相关的功能是受限制的。ESP可以支持加密功能及数据认证和完整性功能的任意组

合，这意味着它可以同时执行两个功能组或仅执行加密功能组。

AH和ESP（表24-1）间的安全功能分配是以限制对加密工具的导入、导出或两者皆有为依据的，这是为很多国家所采用的惯例。这些协议可以单独使用或一起使用。当由于存在的限制不能使用加密时，这个惯例就显得很方便。此时，只能提供系统以AH协议。无疑，在很多情况下仅有AH的数据保护是不够的。在这些情况下，接收端只能检查出数据是否是从所期望的节点发送出来的及数据的格式是否与发送时的一致。当信息通过网络传输时，AH协议不能阻止未授权的察看。这是因为AH协议不对数据进行加密。要对数据加密，就必须使用ESP协议。

表24-1　IPSec协议间的功能分配

功　能	协　议	
保证完整性	AH	ESP
保证真实性		
保证机密性（加密）		
私钥分配		IKE

24.2.3　IPSec中的加密

为了给数据加密，IPSec可以使用任何对称加密算法。在对称加密方法（symmetric encryption methods）中，机密性是建立在发送端和接收端有一个只有他们知道的加密函数的参数的基础上的，这个参数叫做私钥。私钥既用于加密也用于解密。

图24-2阐释了一个对称加密系统的经典模型。这个系统的理论基础由Claud Shannon在1949年最先提出，在这个模型中，有三个参与者：发送端、接收端和入侵者。发送端的目标是使用一个公共信道传送一个受保护的报文。为了这个目的，发送端使用密钥k对明文X加密，之后，发送端传送密文Y。接收端的目标是解密Y以阅读消息X。假设发送端有他自己的密钥提供商。事先产生的密钥经由可靠的受保护的信道传送给接收端。

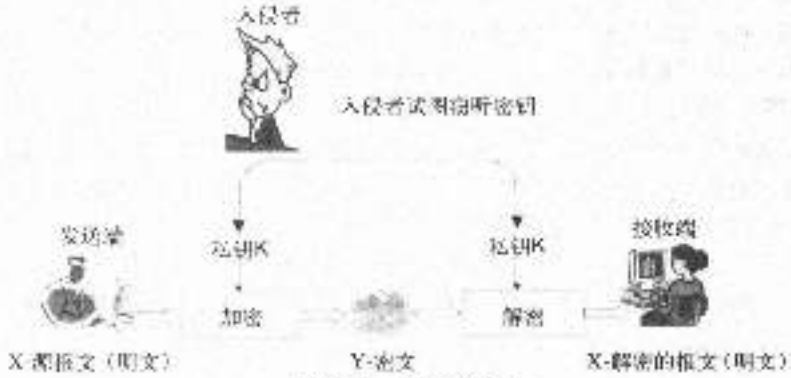


图24-2　对称加密方法

说明　IBM于1976年开发的**数据加密标准**（Data Encryption Standard，DES）在很长一段时间里是最流行的对称加密算法。在2001年，DES被一个更新的更高级的标准，**高级加密标准**（Advanced Encryption Standard，AES）取代，AES提供了安全和性能的一个更佳结合。

保证数据完整性和认证的方法是基于一种使用**单向函数**（one-way function，OWF）的加密技术，这种单向函数也被称为**哈希**（hash）或**摘要函数**（digest function）。

应用于必须加密的数据的这种函数，提供了所谓的摘要值作为结果。OWF必须满足这样一个条件：根据这样函数计算出来的摘要是不能还原源数据的。

在考虑解释摘要函数的使用图例前，注意无论源数据的大小是多少，这种函数产生的摘要值

总是由一个很小的固定数目的字节组成。

假设当传递信息通过不可靠网络时，必须保证数据完整性。为了实现这个目标，所传输数据的摘要必须在发送端计算。

然后摘要和源报文一起在网络上传送（图24-3）。知道产生摘要所使用的OWF的接收端根据收到的报文重新计算摘要。如果从网络上收到的摘要值与本地计算的摘要值相匹配，则源报文在传输期间没有被改变。

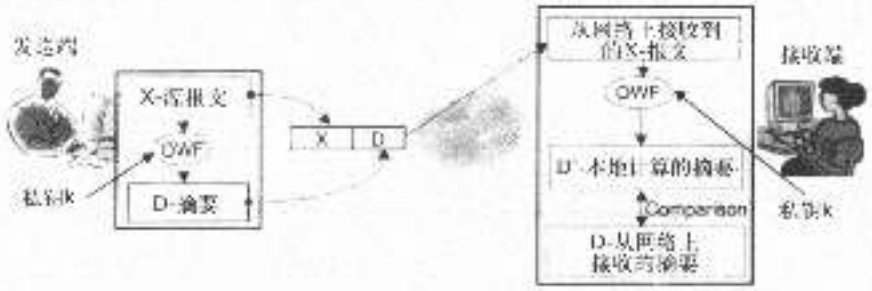


图24-3 使用单向加密来保证数据完整性和真实性

说明 在安全系统中最流行的使用最广泛的摘要函数是MD2、MD4和MD5系列。所有这些函数产生固定长度的摘要：16字节。一个称为SHA的美国标准类似MD4的一个被采用的版本，在这个标准中，摘要长度是20字节。IBM支持MDC2和MDC4摘要函数。

摘要是源报文的一种校验和。然而摘要和校验和之间有很大的区别。校验和用于当传送报文通过不可靠链路时检查消息完整性。它的使用不是针对恶意行为的保护。所传输分组中的校验和的存在不能阻止入侵者增加一个新校验和来欺骗报文。与校验和相比，摘要是使用私钥计算的。如果使用一个带一个参数（扮演只有发送端和接收端知道的私钥的角色）的OWF来计算摘要，那么任何修改源报文的企图都将立刻被发现。

因此，单向函数同时解决了两个问题：它既控制报文的完整性又证明了数据的真实性。这种数据传输方法和其他证明报文发送端真实性的方法，就是ISO术语中的**数字签名**（digital signature）。数字签名的主要应用范围包括电子会议的金融文档和与国际协议有关的文档。最常见的构造数字签名的方法是RSA**非对称算法**（RSA asymmetric algorithm）。这个算法基于Diffie-Hellmann思想。这个思想假设每个网络用户都有一个用于构造加密的数字签名的私钥；所有其他用户使用与这个私钥相对应的公钥来检查签名。

24.2.4 安全关联

为了确保AH和ESP能够保护所传输的数据，IKE协议在两个端点之间建立了一个逻辑连接；在IPSec标准中，这叫做**安全关联**（security association，SA）。

IPSec标准允许受保护的信道的端节点为经由这个信道进行通信的所有主机的流量传输使用一个SA，并允许为了这个目的创建任意数目的SA。例如，每个TCP连接就有一个SA。这保证了选择所需安全等级的可能性，从为多个端节点流量服务的一个普通关联，到使用一个定制SA对每个应用程序的保护。

SA是一个单向的（单一的）逻辑连接；所以，如果需要保证一个安全的双向数据交换，则需要建立两个SA。一般而言，这些SA可以有不同的特征，例如，当仅在一个方向上传送数据库查询时，认证就足够了。另一方面，对于作为响应发送的机密数据，还需要确保机密性。

建立SA的程序从双方的相互认证开始。这是因为如果传送的或接收的数据是来自第三方的话，那么所有的安全机制都将毫无意义。随后要选择SA参数决定AH或ESP协议中的哪一个将用于数据

保护，该安全协议将执行哪些功能。例如，它可以只检查真实性和完整性，或者它还可以确保机密性。SA中的其他重要参数是AH和ESP协议使用的私钥。

IPSec系统提供了自动和手动建立SA的可能性。当使用手动方法时，网络管理员配置端节点以保证它们支持包括私钥在内的协商的关联参数。当使用自动程序建立SA时，运行于信道两端的IKE协议在协商进程期间选择参数。对于AH和ESP协议执行的每个任务，都可以从多个认证和加密协议中挑出（图24-4）。这个能力使IPSec成为一个灵活的工具。注意，为解决保证完整性和真实性问题的而对摘要函数的选择，不影响为保证数据机密性而对加密函数的选择。

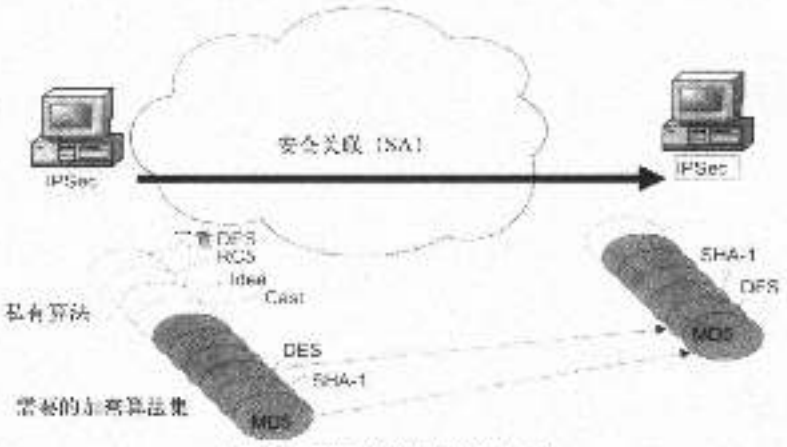


图24-4　ESP协议中的参数协商

为了保证兼容性，IPSec的标准版本定义了一个所需的“工具箱”。例如，数据认证通常需要一个标准的单向加密函数，要么MD5要么SHA-1，并且加密算法的清单上必须包括DES。IPSec产品的制造商通过包含其他认证和加密算法扩展该协议。例如，很多IPSec实现支持流行的三重DES加密算法及一些相关的更新算法，如Blowfish、Cast、CDMF、Idea和RC5。

24.2.5　运输和隧道模式

AH和ESP算法可以以两种模式保护数据：运输和隧道。在**运输模式**（transport mode）下，IP分组使用该分组的原有头部通过网络传送。在**隧道模式**（tunnel mode）下，原分组被封装在新的IP分组，数据是根据新IP分组的头部在网络上传送。

具体模式的使用取决于数据保护的要求和终止安全信道的节点在网络中扮演的角色。例如，这个节点可能是一个主机（端节点）或一个网关（中间节点）。相应地，有下列三种IPSec应用模式：

- 主机-主机
- 网关-网关
- 主机-网关

在第一个模式下，安全信道或SA（在上下文中是一样的）建立于网络两端节点之间（图24-5）。在这种情况下IPSec运行于端节点上并保护从主机1传送给主机2的数据。对主机-主机模式，使用最多的是运输保护模式。

根据网关-网关模式，受保护的信道是建立于两个被称为**安全网关**（security gateways，SG）的中间节点之间的，这两个节点都运行IPSec（图24-6）安全数据交换可以发生在与SG后面网络连接的任意两个端节点之间。端节点不要求支持IPSec。它们以明文方式通过公司的可靠的企业内部网传送它们的流量，送往公共网络的流量经过SG，由它使用IPSec确保流量的保护。网关只能使用隧道模式操作。



图24-5 根据主机——主机模式建立安全信道

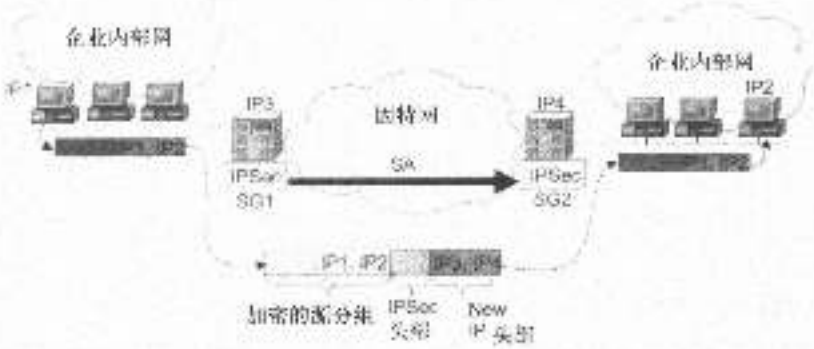


图24-6 根据网关—网关模式的隧道模式下的安全信道操作

如图24-6所示，地址为IP1的用户计算机使用IPSec的隧道模式发送一个分组给地址IP2。SG1网关对整个分组加密，包括IP头部，并提供被加密分组一个新头部。在新头部中，它指定它的地址IP3作为源地址，并填充目的地址IP4为SG2的地址。所有的数据是基于扩展分组的头部的数据在IP上传送。封装的分组用作扩展分组的数据域。当分组到达SG2网关后，IPSec获取封装的分组并将它解密，从而把它还原为它的初始格式。

主机—网关模式（图24-7）通常用于远程访问。在这种情况下，安全信道存在于运行IPSec的主机和保护所有连接到公司企业内部网的主机的网关之间建立一条安全信道。如果你在远程主机和属于该网关保护的内网中的任意其他一台主机间另建一个安全信道，则这个模式会变得更加复杂。两个SA的这种组合使用将可靠地保护内网中的通信。

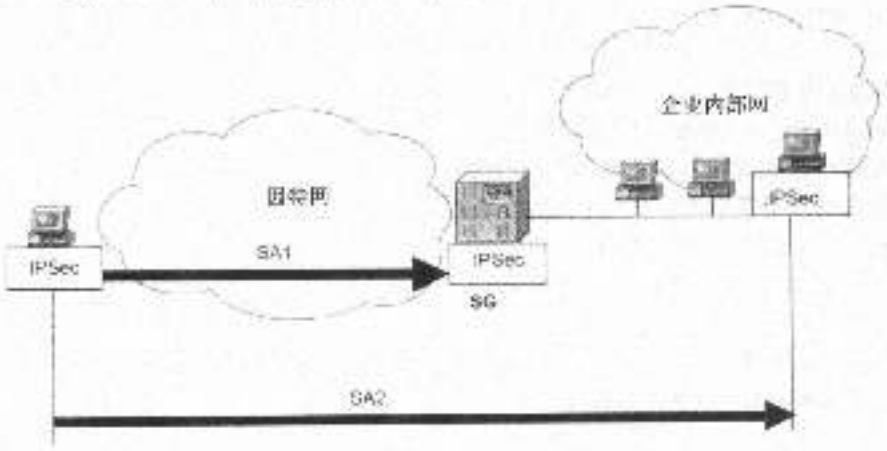


图24-7 安全的主机—网关信道设计

24.2.6 AH协议

AH协议允许接收端保证下列几点：

- 一方将分组发往已建立的关联
- 在通过网络传输期间分组内容不被修改
- 分组不是前面已接收的某些分组的副本

前面两个功能是对AH协议的要求。第三个功能是可选的，可以在建立关联时选择。为了执行这些功能，AH协议使用了一个特殊的头部。该头部的格式如图24-8所示。

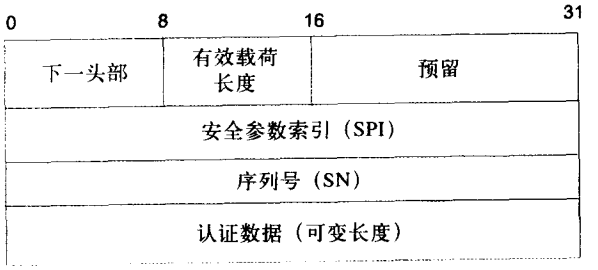


图24-8　AH协议头部的结构

*下一头部*字段指定高层协议的编码——例如，报文封装在IP分组数据字段的协议。这可能是某个运输层协议（TCP或UDP）或ICMP。然而，这也可能是ESP协议，如果它与AH一起使用的话。

*有效载荷长度*字段指定了AH头部的长度。

下一个字段，*安全参数索引*（*SPI*），用于关联分组和它对应的SA。

序列号（*SN*）字段指定分组的顺序值。它用于保护分组免于被第三方复制，该第三方试图再利用探测到的认证用户发送的受保护分组。发送端有序地增加每个在该关联的框架内传送的新分组中该字段的值，因此副本的到达将很快被接收端发现——如果在该关联框架内针对错误副本的保护是可行的。无论如何，AH协议都不会恢复丢失的分组也不会对到达的分组重新排序。如果它发现接收到一个相似分组，那么它将丢弃该副本。为了减少协议操作所需的缓存大小，使用了滑动窗口算法。只对那些序列号落在窗口范围内的分组检查副本。通常，窗口大小设置为32或64个分组。

*认证数据*字段用于对分组进行认证及检查完整性。这个字段包含**完整性检查值**（Integrity Check Value，ICV）。这个值，也被称为摘要，是使用某个AH协议必须支持的OWF加密函数——MD5或SAH-1计算得来的。然而，它也可以使用任意一个双方在建立关联时一致同意的可选函数。当计算摘要时，对称私钥被当作一个参数。关联的私钥可以手动指定也可以使用IKE协议自动设定。因为摘要的长度依赖于所选择的函数，所以通常这个字段有一个可变长度。

当计算摘要时，AH协议尽量考虑尽可能多的源IP分组字段。然而，这些字段中的某些字段会在分组通过网络传输期间发生不可预料的改变。所以，这些字段不能包括在分组的认证部分中。例如，在接收节点不能用生存时间来评价分组的完整性，因为这个字段的值每经过一个中间路由器都会被减1，所以它必与源值不同。

分组中AH头部的位置取决于安全信道配置的模式——运输或隧道。图24-9阐释了在运输模式下，最终的分组是以何面貌出现的。

如果使用的是隧道模式，当IPSec网关接收到中间分组并把它封装进扩展IP分组中时，AH协议既保护外部分组头部的末改变字段也保护源分组的所有字段（图24-10）。

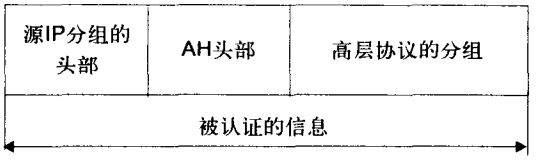


图24-9 在运输模式下AH协议处理的IP分组的结构

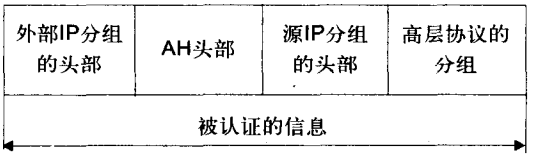


图24-10 在隧道模式下AH协议处理的IP分组的结构

24.2.7 ESP协议

ESP协议解决了两组问题。第一组包括与AH协议类似的功能，即，在摘要的基础上保证数据完整性和认证。第二组包括通过对所传数据加密来防止未授权探察的数据保护。

如图24-11所示，头部被有效载荷数据字段分为两部分。第一部分，*ESP头部*，由两个字段组成，*SPI*和*SN*，它们与同名的AH字段类似。头部位于数据字段前面。剩下的ESP协议服务字段叫做*ESP标尾*，位于分组末端。

这两个标尾字段类似于AH头部的两个字段，它们是下一头部和认证数据字段。如果ESP在建立SA时没有使用与保证数据完整性相关的功能，则认证数据字段缺失。除了这两个字段外，尾部还包括两个辅助字段——填充和填充长度。在三种情况下可能会需要填充。第一，一些加密算法的普通操作要求要加密的明文包含偶数个预定义了大小的块。第二，ESP头部要求数据字段终止于4字节边界。最后，填充可以用来隐藏数据字段与大小以保证所谓的流量部分的机密性。隐藏能力受相对较小的255字节的填充大小限制。这是因为大量的辅助数据会减少通信链路的有效带宽。

图24-11展示了运输模式下ESP头部字段的位置。在这个模式下，ESP不对IP分组头部加密；如果它这么做了，路由器将不能识别头部字段和在网络间正确转发分组。必须加密的字段清单中也不包含SPI和SN，它们必须以明文方式传递，以保证到达的分组可以按所属的关联分类。这保护分组免遭未授权复制。

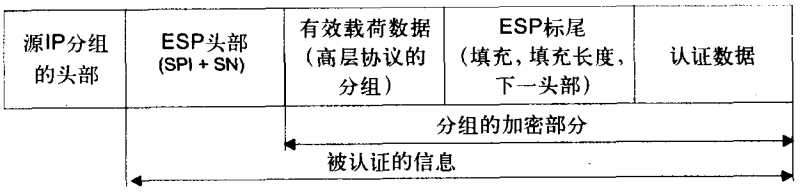


图24-11 运输模式下ESP协议处理的IP分组的结构

在隧道模式下，源IP分组头部被置于ESP头部之后，并落在保护字段的清单中。外部IP分组的头部不受ESP协议的保护（图24-12）。

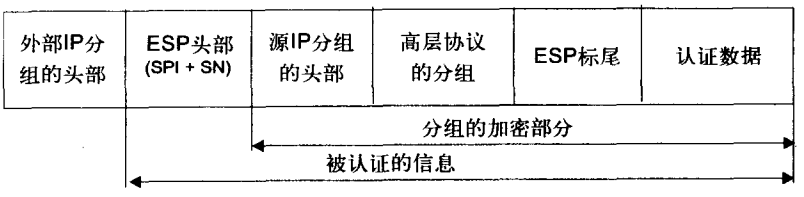


图24-12 隧道模式下ESP协议处理的IP分组的结构

24.2.8 安全数据库

因此，IPSec提供了多种流量保护方法。那么运行于主机或网关的IPSec实现，如何选择应用于流量的保护方法？解决方法是使用每个节点支持的IPSec的两种数据库：

- 安全关联数据库（SAD）
- 安全策略数据库（SPD）

当建立SA时，和建立其他逻辑连接一样，双方缔结多个调节相互之间流量传输的协议。协议以参数集形式创建。对SA，这些参数是安全协议（AH或ESP）的类型、协议运行模式、加密方法、私钥、当前关联中当前分组的序号和其他信息。决定所有活动关联的参数集以SAD形式存储在受保护信道的两个端节点上。每个IPSec节点都支持两个SAD，一个用于入关联，一个用于出关联。

另一种类型的数据库是SPD，它指定IP分组间的映射并为它们建立的规则进行处理。SPD记录由两种字段组成——分组选择器字段以及具有当前选择器值的分组的安全策略字段（图24-13）。

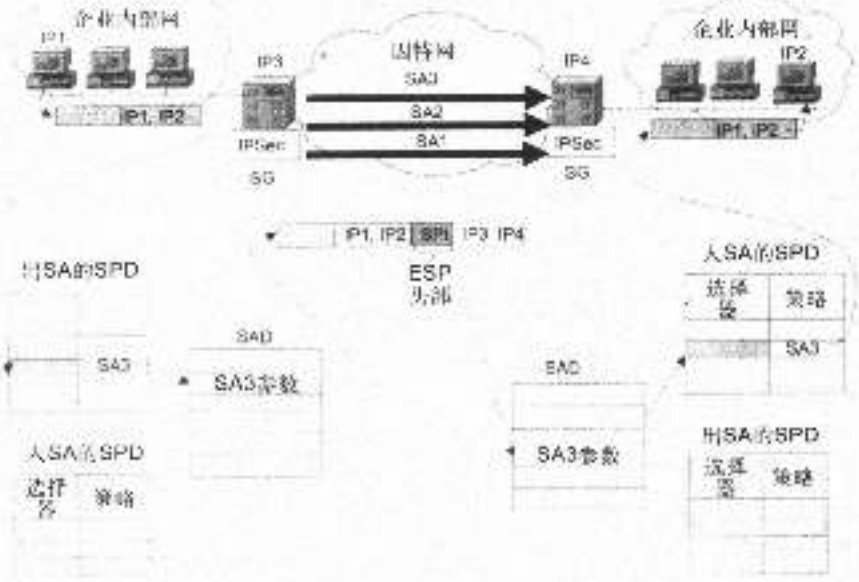


图24-13　使用SPD和SAD

SPD内的选择器包含下列属性集，它们使得SPD可以检测需要保护的流量：

- 源及目的IP地址，它们可以由一个任意类型的（单独的、组、或广播的）单独的地址或由上下限或使用一个带掩码的地址指定的地址范围表示
- 源及目的端口（如，TCP或UDP端口）
- 运输层协议的类型（TCP，UDP）
- 用DNS或X.500格式表示的用户名
- 用DNS或X.500格式表示的系统名（主机、SG、等）

对每个到达安全信道的新分组，IPSec检查SPD中所有的数据库记录，并将选择器值与IP分组对应字段相比较。如果该字段的值与指定的选择器相匹配，且对该分组执行定义在该记录安全策略字段的操作。策略提供下列可能性中的一个：不做修改传送分组、丢弃分组或使用IPSec处理分组。

在后一种情况下，安全策略字段必须有时包含该分组SA参数集的SAD记录。（在图24-13所示的例子中，SA3是为该出分组定义的）。基于为SA3指定的参数，将某个协议（图24-13中为ESP）应用于该分组、加密函数和私钥。

如果必须对这个出分组应用某个安全策略，但SPD记录指示没有所需策略的活动的SA，则IPSec就使用IKE协议创建一个新关联。在这种情况下，新记录将被插入到SAD和SPD中。

SPD可以由用户手动创建和管理（更适合于主机的变体），也可以由系统管理员（更适合网关的变体）或由应用程序自动创建和管理。

在本章前面，我们解释了出IP分组和为它指定的SA之间的关系是如何建立的。然而，这里还

有另外一个问题：接收IPSec的节点如何决定怎么处理到达的分组。毕竟，当使用加密时，很多组成选择器的关键参数将变得不可用；从而，对应的SA参数也将不可用。为了解决这个问题，AH头部和ESP头部提供了前面提到过的*SPI*字段。这个字段包含了一个指向存储对应SA参数的SAD行的指针。这个字段由AH和ESP协议在安全信道起始处处理该分组时填充。当分组到达安全信道的端节点时，从它的ESP或AH头部（在图24-13中，为ESP头部）回收SPI指针。所有的进一步处理都需要考虑该关联指针指示的所有参数。

因此，为了识别对应不同的SA的分组，使用了下列方法：

- 在发送端节点——选择器
- 在接收端节点——SPI

在对分组解密后，接收IPSec的节点检查它的属性（现在可用的那些）以匹配入流量的SPD。这既保证了这里不会有错也保证了对应于管理员指定的安全策略的分组处理。

使用SPD和SAD控制流量保护，给面向连接的SA机制和IP流量的数据报性质的结合带来了灵活性。

24.3 虚拟专用网服务

24.3.1 VPN定义

术语虚拟专用网（VPN）指的是这样的网络，它们具有真正专用网的特征。网络只有在被公司所有且公司拥有对所有网络基础设施——电缆、交叉设备、信道-构建设备、交换机、路由器和其他通信设备的完全控制时，才可以被认为是专用。

专用网区别于共享网络或公共网络的主要特征是它和其他网络的**隔离**（isolation）。

这些隔离的结果如下：

- 独立的网络技术的选择——唯一受限制的就是销售商或制造商的选择
- 独立的寻址系统——在专用网中，可以选择任意地址
- 可预测的性能——拥有通信链路，保障了公司端节点之间（对于WAN链路）或通信设备之间（对于本地连接）的预定义的带宽
- 最大程度安全性——缺少与外界的连接很大程度上减少了来自外网攻击的可能性。它也减少了路径上对流量窃听的可能性。

然而，专用网是一个极不经济的解决方法。只有大公司才能支付起这种网络，尤其是全国或国际范围的专用网。创建专用网的奢华是那些拥有创建专用网所需的全部基础设施的用户的特权。例如，大天然气或石油公司可以沿着它们的管道安装专用的技术电缆系统。在公共数据网络的基础设施达到所需的发展程度前，专用网络直到最近才流行起来。如今，几乎所有的这种网络都被VPN排挤出了市场，VPN提供了QoS和它们提供的服务代价之间的一个折衷。

VPN技术允许使用一个被多家公司共享的介质来实现一些服务，这些服务的QoS特征可以与专用网相媲美（包括安全性、可用性、可预测带宽及独立的地址系统的选择）。

根据实现它们的对象，VPN可以分为两类：

- **用户提供的VPN**（customer-provided VPN，CPVPN）这个名字反映了所有与VPN支持相关的问题都必须由用户自己解决。在这种情况下，ISP仅提供访问公共网络以连接用户的端节点这个传统服务。公司雇用的网络专家配置并管理VPN工具。
- **提供商供应的VPN**（provider-provisioned，PPVPN）这个术语反映了服务提供商使用自己的网络为每个用户构造一个专用网这个事实。“专用”用户网络与其他网络隔离并被保护起

来。这个VPN组织方法相对更新一点。因此，它没有第一种方法应用广泛。

在最近几年，PPVPN的流行度稳步提升，因为创建和支持VPN这项工作困难且特殊。从而，大多数公司喜欢把这项工作交给一个可靠的服务提供商。VPN服务的实现允许服务提供商提供用户一定范围的额外服务：对网络操作的控制、Web和电子邮件服务管理、及专用软件管理。

为了可以把VPN分为CPVPN和PPVPN外，还有另外一种分类。根据这种分类，VPN可以由执行VPN功能的设备的位置来划分。VPN可以分为：

- 基于安装在用户驻地设备的：**基于用户驻地设备的**（customer premises' equipment-based，CPE-based）或**基于用户边缘的**（customer edge-based，CE-based）VPN。
- 基于服务提供商的基础设施的：**基于网络的**VPN（network-based VPN）或**基于提供商边缘的**（provider edge-based，PE-based）VPN。

在VPN类型名称中的术语边缘说明了大多数（有时是所有）与VPN支持相关的功能都是由用户或服务提供商拥有的边缘网络设备执行。

提供商支持的网络可以属于PE-based类型或CE-based类型。第一个变体更常见些，因为提供商维持连接到自己网络的设备。在第二种情况，VPN设备位于用户领域，但是服务提供商远程控制该设备。这使得用户雇员可以从相当困难的任务中解救出来。

当VPN是由用户支持时（CPVPN），所有的设备通常都位于用户网络。这是CE-based类型VPN。

24.3.2　VPN评价和比较的准则

VPN，和其他仿真系统一样[⊖]，第一特征是被仿真对象的特征，第二特征是与原对象的近似程度，第三特征是使用的仿真工具。

因此，考虑在VPN中仿真的专用网元素。

实际上，所有的VPN都模拟提供商基础设施中的为多个用户服务的专用的、租用线路。

说明　这里可能会有一些技术上的困惑。之所以会这样是因为在通信运营商基础设施中在TDM（电话、PDH或SDH）基础上模拟的私有、专用链路通常不被认为是VPN。通常基于用户拥有的网络设备，但是使用的是租用物理链路构建的网络被认为是专用网。这是因为在这些网络中使用的同步TDM技术保证了不同用户的信息流量是独立的。这个技术也保证了每个流量都有一固定带宽并保证了共享介质级的QoS参数。因此，这些链路被称做专用的——它们是为用户个体专门使用而提供的，其他用户不能使用它们。这个方法也为大多数专用网所使用，它是基于租用线路而不是专用线路运行的。

当仿真同一家公司拥有的链路的基础设施时，VPN服务被称为**企业内部网**（intranet）服务。当这些信道是由连接用户和合作公司的链路实现的，并且它们之间的数据交换在安全模式下进行时，这个服务就叫做**企业外部网**（extranet）。

只有当“专用的”物理链路是由ATM、帧中继、X.25、IP或IP/MPLS这样的分组交换技术仿真的时候，才使用术语VPN。在这种情况下，由这些VPN提供的通信质量和那些由**真正的**（actually）专用物理链路提供的通信质量之间的差别是明显的。实际上，虚拟一词的使用之所以合理是因为带宽的不确定性及其他一些开始显现的特征。当使用分组交换网络构建VPN时，用户不仅被提供物理链路，还被提供某个数据链路层技术（如ATM或帧中继）。当使用IP时，用户也被提供网络层服务。

VPN也可以仿真高层网络元素。例如，VPN可以通过支持用户IP流量，并同时创造一个独立的IP网络的效果来运行于网络层。在这种情况下，除了对物理链路的仿真外，VPN还需对用户流

⊖　在这个例子中，VPN被认为是对某公司的专用网的仿真

量执行一些附加操作，包括收集多种统计数据并过滤和反映出用户之间及同一家公司的各子公司之间的相互作用（这种特点不能与和外部用户相独立而混淆，后者是VPN的主要功能）。

在VPN中，与运输服务的仿真相比，应用层服务的仿真是很少见的。然而，它也是有可能的。例如，服务提供商可以支持客户的Web站点，电子邮件系统或专门的资源规划应用程序。

当比较VPN时，使用的另一个准则是，比较VPN服务与真正的专用网提供的服务的近似程度。

安全性是专用网服务的最重要特征。VPN安全性包括受保护的网络整个属性集——信息机密性、完整性、在公共网上数据传输期间的可靠性、为了用户和提供商网络的内部资源免遭外部攻击所提供的保护。VPN安全级别可以根据使用的安全工具——流量加密、用户和设备认证、地址空间隔离（例如，基于NAT）、虚拟信道、点到点隧道和复杂的未授权用户连接，而有很大幅度的变动。然而，因为没有单独的保护机制提供保障，所以安全工具可以结合使用以创造深入的保护。

第二，保证VPN服务接近于真正专用网*QoS*特性是很理想的事。首先，运输服务质量对用户流量采用了有保障带宽。这可能是由其他QoS参数实现的，如最大延迟和可以容忍的数据丢失率。在分组交换网络中，流量突发、可变延迟和分组丢失是不可避免的；所以，虚拟信道与TDM信道的近似程度通常都是不完全的，并且具有概率论的性质。平均起来，对单个的分组是没有保障的。不同的分组交换技术保证的QoS级别不同。例如，在ATM技术中，QoS机制是最理想的并且测试结果也很好。在IP网络中只是在最近才出现类似的机制。所以，不是每个VPN都能仿真出专用网的这些特征。因此，安全性是任何VPN的强制性特征，而它的运输服务质量虽然很理想，但仅是一个可选特征。

第三，如果VPN保证提供给用户独立的地址空间，则它将可以与真正的专用网媲美。这给用户带来了网络配置的便利和维持安全的方法。同时，保证用户对其他用户的地址空间以及提供商主干地址空间一无所知也是很理想的。在这种情况下，提供商主干就可以可靠地防止恶意行为和人为错误。从而，VPN将保证更高质量的服务。

用于创建VPN的技术对VPN特征有相当大的影响。根据保证数据传输安全性的方法， VPN技术可以分为如下两类：

- 流量分离技术
- 加密技术

24.3.3 在流量分离基础上的VPN

流量分离技术使用永久虚电路技术来保证对每个客户流量的可靠保护，防止其他公共网络用户有意或无意的访问。这种类型的技术包括：

- ATM VPN
- 帧中继VPN
- MPLS VPN

点到点虚电路模拟专用线路服务的方式为，经由提供商网络从某个**客户站点**（client site）的CE设备传递到其他客户站点的CE。

说明 在这种情况下，术语站点指示的是客户网络的一个独立片断。例如，连接总部和三个远程子公司的公司网络是由四个站点组成的。

数据保护是通过保证未授权用户在没有修改提供商的设备上的交换表的情况下，不能连接到虚电路来实现的。这阻止了未授权用户执行攻击或读数据。流量保护是虚电路技术的一个固有特征；因此，ATM VPN和帧中继VPN服务仅仅是ATM或帧中继网络的普通PVC服务而已。每个使用PVC基础设施连接LAN的ATM或帧中继网络用户都使用PVC服务。与数据报技术相比，这是虚电路技术普通优点中的一个，因为如果不使用额外的VPN工具，数据报技术不提供用户任何防止其

他用户攻击的保护。

因为ATM和帧中继网络在数据传输期间都只使用两个线等级，所以在它们基础上建立的VPN变体被称为第2层VPN（L2VPN）。在ATM和帧中继中，QoS支持机制的可用性充分考虑了对专用线路质量的良好模拟。

这种网络从来不会对第3层信息进行分析或修改。这既是这种网络的优点也是它们的缺点。优点是客户可以使用这些虚拟信道来传送任何协议的流量，而不仅是IP的。此外，客户和提供商的IP地址是隔离且独立的，它们可以任意选择，因为它们不用于经提供商主干的流量传输。除了虚拟信道标签外，客户不需要有任何关于提供商网络的信息。这种方法的缺点是，提供商不能对客户IP流量进行操作，从而，不能提供客户与IP有关的额外服务。注意，如今这些额外服务是一个吸引了众多提供商的相当有前景的商业领域。

L2VPN的复杂性很高。它的代价也一样，当组织一个全连通的客户站点拓扑时，配置操作对站点数量的依赖是一个二次关系（图24-14a）。

为了连接N个站，需要创建$N\times(N-1)/2$个双向虚电路（或$N\times(N-1)$个单向虚电路）。特别地，如果$N=100$，则需要5 000个配置操作。虽然这些操作是由网络管理系统自动执行的，但是，还是有手动操作和错误的可能。当只提供企业内部间服务时，连接要求的配置数直接与客户成比例，这当然是一个优势！企业外部网的使用使这个情况复杂化，因为在这种情况下需要保证不同客户站点间的连通性。如果客户放弃全连通拓扑并通过一个或多个专用传输站来组织星型拓扑，则ATM和帧中继VPN的可延拓性会得到改进（图24-14b）。自然，客户网络的性能会有所下降，因为将会有更多的转发信息传输。然而，这种方案的经济性是很明显的，因为通常提供商都是根据信道来对客户收费的。

ATM或帧中继VPN的客户不能对彼此构成威胁。此外，他们也不能攻击提供商网络。如今，任何一个提供商都拥有一个IP网络，即使它只提供ATM或帧中继VPN服务。限制自己只提供ATM或帧中继服务的提供商是很少见的。毕竟，如果没有IP网，提供商将不能管理和维持它的ATM或帧中继网络。所以，ATM或帧中继会对它的结构甚至它的存在一无所知，见图24-15所示。

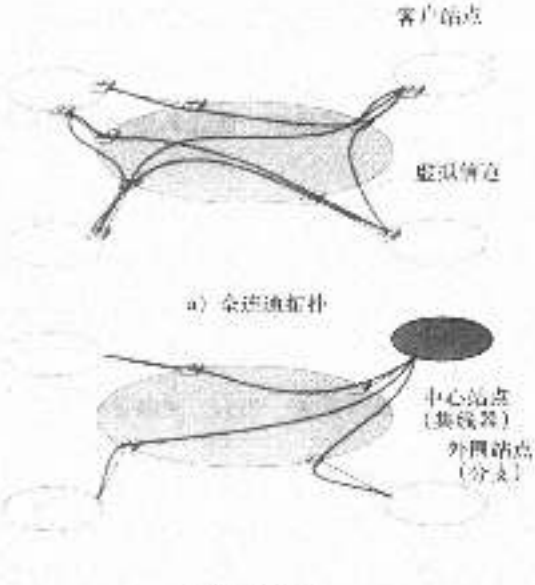


图24-14　L2 VPN可延拓性

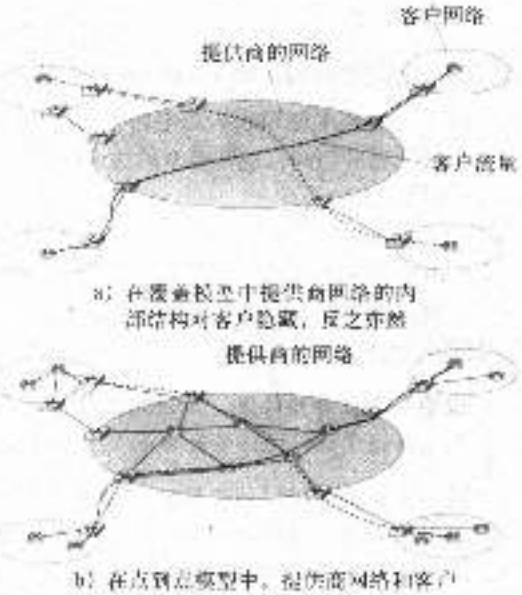


图24-15　覆盖VPN模型和点到点VPN模型

*MPLS VPN*被分为MPLS VPN 2L（2层）和MPLS VPN 3L（3层）。这两个技术都在MPLS提供商网络中为客户流量分离使用标记交换路径（LSP）。

MPLS VPN 3L根据IP地址与客户网络通信。出于同一个目的，MPLS VPN 2L使用2层地址信息。例如，这些信息可能是MAC地址或帧中继虚电路的标识符。

使用MPLS网络能为服务提供商极大地减少与VPN配置有关的手动操作。与MPLS VPN配置相关的操作数和客户站点数目成正比。注意，对于ATM或帧中继VPN，这个值是与客户站点数目的平方成比例。MPLS VPN的另一个优点是VPN和其他MPLS应用的紧密结合，这些应用有TE和QoS。因为对于MPLS VPN的特殊意义，这个话题将在本章后面具体阐述。

24.3.4 IPSec VPN

基于**加密**（encryption）的技术是另一类VPN技术。它们用于在数据报网络的基础上构建VPN，数据报网络不能保证流量分离。经典的IP网络就属于这一类型。

IPSec是主要的基于加密的VPN技术。它用于创建一个连接属于一家（企业内部网）或多家（企业外部网）公司的多个站点的安全信道的基础设施。

IPSec标准提供高度灵活性，它不但允许公司在多种认证和加密算法中选择所需要的，还允许选择所需要的保护模式（加密或保证数据真实性和完整性）。IPSec封装模式通过使用两个IP地址——外部和内部的，隔离了客户和服务提供商的地址空间。

IPSec最常用于支持CPVPN，在这种情况下，客户创建经过提供商网络的IPSec隧道。对提供商的要求仅是提供标准的网络互联服务。从而，客户既可以访问提供商网络内的可用服务也可以访问因特网服务。IPSec VPN的配置程序是很复杂的，因为IPSec隧道是点到点隧道。当实现的是全连通拓扑时，这些隧道的数量与$N\times(N-1)$成比例。它还需要考虑一个支持关键基础设施的艰难任务。

IPSec也可以用于创建提供商支持的VPN。在这种情况下，隧道再次建立在CE-based设备基础上；然而，这些设备是提供商远程配置和维持的。

在真正的专用网的所有特征中，IPSec VPN仅模拟安全性和对地址空间的隔离。

这个技术不支持链路带宽和其他QoS参数。然而，如果服务提供商保证QoS服务（例如，使用DiffServ），当创建IPSec隧道时也可以使用这些服务。

基于加密的VPN技术可以和基于流量分离的VPN技术一起使用，以提高它们的安全级别。因为安全级别的不足，基于流量分离的VPN技术经常成为被批评的对象。有时，客户认为缺少流量加密会允许提供商私下未授权数据访问。这种可能性确实存在；因此，接收如MPLS VPN这样的基于流量分离的VPN服务的用户，可以加强流量安全性，例如通过使用IPSec虚电路技术。

24.4 MPLS VPN

如今MPLS VPN吸引了普遍的关注。提供商提供客户的服务范围也在不断地增长。这使得MPLS VPN可以被全世界众多用户使用。与包括ATM、帧中继和IPSec在内的其他VPN构建方法相比，MPLS VPN看上去更有优势。这是因为它的高可延拓性、自动配置的可能性、以及可以与任何提供商提供的其他成功的IP服务的自然结合，这些IP服务有：因特网访问、Web和邮件服务及托管。

MPLS VPN有以下两种：

- MPLS L3VPN，在MPLS L3VPN中，从客户到提供商网络边界设备的流量传递是使用IP技术（3层）执行的。
- MPLS L2VPN，在MPLS L2VPN中，客户流量是使用任意一个2层技术传送到提供商网络中

的，这些2层技术如，以太网、帧中继或ATM。

在这两种情况下，提供商网络内客户流量的传输都是使用MPLS技术执行的[6]。

本书中，只详细提及了MPLS L3VPN，因为它是已被很多提供商网络使用的最成熟的并且是可以信赖的技术。虽然定义该技术的主要机制的RFC2574bis规范只有信息地位，但是所有的MPLS L3VPN实现都是依据此文件，因此这使得它成为实际上的标准。在进一步的讨论中，符号L3 将被省略，术语MPLS VPN将等同于MPLS L3VPN。

24.4.1　完全连接和绝对隔离

每个客户都希望VPN服务提供商连接他的网络，并确保最终互联网独立于其他客户的互联网。

当代的提供商必须在用作通用运输的IP技术的支配下解决这个问题。IP互联网操作的一个主要原则是所有网络自动互联为统一的互联网。这是通过使用多种路由协议如BGP、OSPF和IS-IS在整个互联网上传播路由信息来实现的。这个方法允许在每个网络路由器上自动创建一张路由表。路由表指定分组传递给每个分组网络所要经过的路径。虽然到某些网络的路径可能会聚合，但这没有改变这个问题的本质。

MPLS VPN技术如何解决保持连通性的同时保证隔离这个矛盾？它的解决方法十分高明：自动过滤路由广告并使用MPLS隧道来传送客户流量通过提供商内部网络。

为了隔离网络，设置一个阻塞路由信息传播的屏障就足够了。为了在网络限制内交换路由信息，节点使用一个内部网关协议（IGP），它的应用范围受自治系统限制，这些系统有：RIP、OSPF或IS-IS。如果节点A的路由表不包含到节点B的路由记录，并且默认路由也没有该记录，那么节点A“看不见”节点B。

在MPLS VPN中，这之所以得以实现，是因为来自客户网络的路由广告使用BGP在整个服务提供商内部网络上“跳跃”。因为一个特殊配置程序，该程序是由一个使用名为多协议BGP（multiprotocol BGP，MP-BGP）的BGP扩展版本执行的，这些路由广告仅被传送给同一个客户的网络。结果，不同客户的路由器相互间没有对方的路由信息。所以，它们不能交换分组，这也意味着实现了想要的隔离（图24-16）。

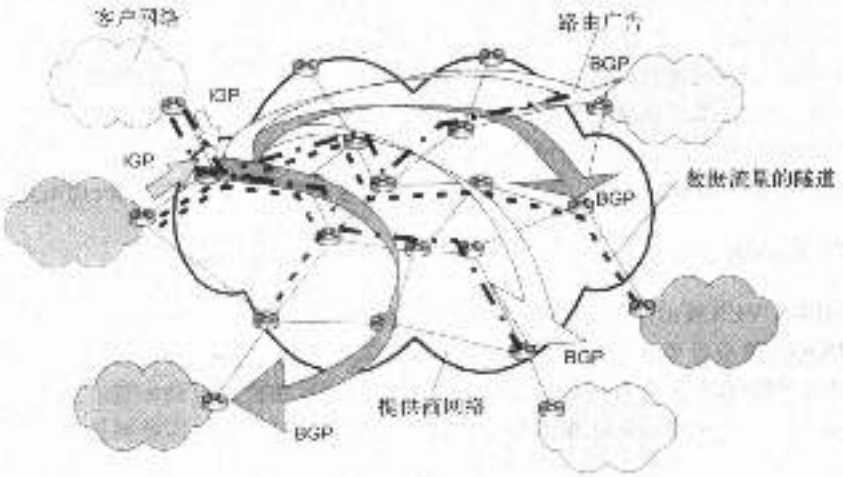


图24-16　使用隧道隔离客户网络

[6] 目前，很难判定MPLS级别，因为这个领域的术语还没有很好地定义。然而，因为基于对应于2层的本地标表的分组转发，本书中MPLS将被定义为2层技术。

这个方法的另一结果是提供商内部网络的隔离，这提高了它的可靠性和可延拓性，因为在这种情况下，没有支持大型路由表的必要，这些大型路由表存储关于服务提供商网络的内部路由器上的众多客户网络的信息。

然而，还有一个需要解决的问题：对于地理上分布的客户网络，如果服务提供商网络在标准路由表中没有它们的信息，那么如何把它们组织为一个统一的VPN？为了实现这个目标，使用了一种传统技术，即，在内网中的边界路由器间创建隧道。目前考虑的技术使用的是MPLS隧道，其他可选的解决方法有：创建IPSec隧道或其他"IP之上的IP"隧道。MPLS VPN的优点在于它的自动创建和配置。另外，MPLS VPN隧道有MPLS的所有一般特性：它保证加速转发（与路由相比）和确保流量工程。在现实世界的网络中，为了实现前面描述的创建MPLS VPN原理，开发了很多专门的方法和网络部件。

24.4.2 MPLS VPN部件

在任何MPLS VPN中，都能发现下面两个主要区域（图24-17）：

- 属于客户的IP网
- 属于提供商的MPLS主干

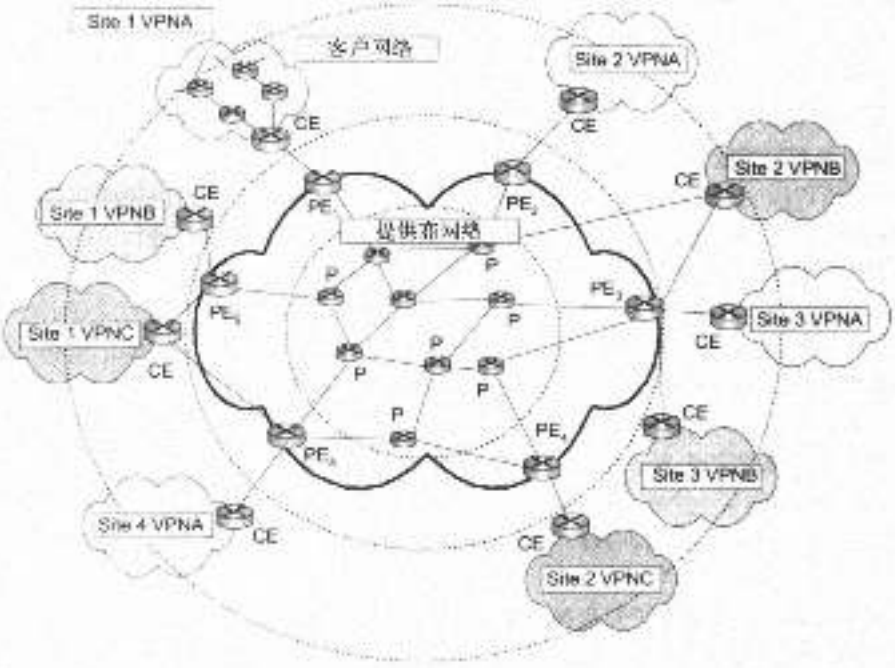


图24-17 MPLS VPN 部件

一般而言，每个客户都有多个地理分布的IP网络（站点）。每个站点可能包含多个通过路由器连接的子网。同一客户拥有的站点在提供商网络上交换IP分组，并组成了该客户的VPN。

CE路由器就是客户站点连接提供商主干所经由的那个路由器。例如，作为客户网络一个部件的节点C，它没有关于VPN存在的任何信息，它可以通过多个链路连接到提供商主干。

提供商主干是一个MPLS网络，IP分组是基于本地标记而不是IP地址转发的。MPLS网络由标记交换路由器（LSRs）组成，它根据标记值引导流量沿着事先创建的LSP传输。

在提供商网络，有两类LSR设备：提供商边缘路由器（PE路由器），客户站点经由CE路由器与它连接；提供商主干的提供商路由器（P路由器）。CE和PE路由器通常是直接由一条物理链路直接

连接的。在该链路上运行着某数据链路层协议，例如PPP、帧中继、ATM或以太网。CE和PE间的所有通信都发生在TCP/IP栈的标准协议基础上。只有PE的内部接口（和所有的P接口）需要MPLS支持。有时把转发流量分为**入**（inbound）PE和**出**（outbound）（**远程**）PE会很有用。

在提供商主干中，为了支持VPN，只有PE路由器是必须要配置的，因为只有它们拥有与已有VPN相关的信息。

如果你从VPN角度来考虑某个网络，那么提供商路由器不与客户CE路由器直接通信。相反，它们位于入PE路由器和出PE路由器间的路径上。

PE路由器的功能比P路由器更复杂。这些路由器执行与VPN支持相关的主要任务，即，路由和不同客户数据的分离。PE路由器也是客户站点间LSP的端节点。PE给赋予经由P路由器内部网络传输的IP分组分配标记。

可以使用两种方法创建LSP：基于LDP的快速IGP路由和带RSVP或CR-LDP的流量工程技术。LSP的创建，在组成当前LSP的所有的PE和P路由器上创建了**标记-交换表**（label-switching table）（第22章提供了这些表的例子）。同时，这些表还指定了用于不同客户流量的路径集合。VPN实现了多种链路拓扑：全连通、星型（通常叫做中心辐射），或网状。

24.4.3　路由信息的分离

为了VPN正确运行，有必要保证使用提供商主干的路由信息的传播不超过它的范围。此外，客户站点内的路由信息也不能被VPN范围的外界获知。

阻止路由广告传播的屏障可以通过适当地配置路由器来实现。路由协议必须从接口处获得信息，这些接口既有权接收广告，也有权传播广告。

在MPLS VPN中，这些屏障的角色被赋给了PE路由器。PE路由器可以用作客户站点域和提供商网络核心域之间的不可见的边界。在边界的一边是PE路由器与P路由器通信所经由的接口；另一边是客户站点所连接的接口。在主干内已有路由的广告到达PE的一边；客户网络内的路由广告从另一边抵达。

图24-18展示了路由信息分界的方法。在PE路由器上安装了多个IGP实体。其中一个实体用于接收和传播路由广告，这些广告仅来自于连接该PE路由器和P路由器的内部接口。另有两个IGP实体处理来自客户站点的路由信息。

其他PE以类似的方法配置。P路由器接收并处理来自所有接口的IGP路由信息。最后，在每个PE和P路由器上都创建了一个路由表。这些路由表包含有关提供商内网的所有路由信息。有必要指出P路由器上的路由表不包含客户网络内已有路由的信息。客户网络没有有关提供商网络内的路由信息。

在每个PE路由器上都创建下列两种路由表：

- **全局路由表**（Global routing table），是在来自主干的广告的基础上创建的
- **VPN路由和转发**（VRF）**表**（VPN routing and forwarding table）——PE在来自客户站点的广告的基础上形成的表

客户站点是普通的IP网络，它可以使用任何IGP路由协议对路由信息进行传递和处理。显然，服务提供商不控制该过程。路由广告在每个站点内自由传播直至它们抵达作为阻止它们进一步传播的屏障PE路由器。

通过在连接客户站点的PE路由器的每个接口上安装一个单独的路由协议，可以保证不同客户的路由的分离。为某接口定义的协议只接收和传送来自该接口的客户路由广告，且它不把这些广告发送给连接PE到P路由器所经过的中间接口。这些广告没有发送给其他客户的站点。结果，在PE路由器上产生了多个VRF路由表。

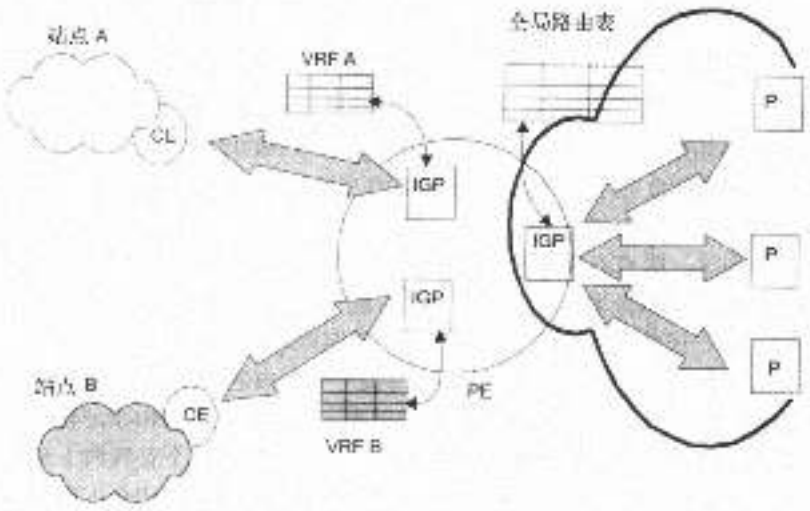


图24-18 路由信息分离方法

精确对该情况进行下简化，可以认为在每个PE上创建了多个VRF表。这些表的数量对应于连接到该PE的站点数目。实际上，在PE路由器上组织有多个**虚拟路由器**（virtual router），每个虚拟路由器与它自己的VRF表一起运行。

站点和VRF表之间可能还有其他关系。例如，如果同一个VPN的多个站点与同一个PE相连，则有可能为它们创建一张共用的VRF表。图24-18展示了两张VRF表：一张包含对站点A的节点路由的描述，另一张包含到站点B的路由信息。

24.4.4 用MP-BGP连接站点

为了将地理分布的客户站点连接成统一的网络，有必要为它们创建一个公共空间以传播路由广告，然后在内网内创建受保护的路径，在这些路径上，属于同一VPN的不同站点的节点可以安全地交换数据。

同一VPN的不同站点交换路由信息的方法是MP-BGP。该协议的详细描述可见RFC 2858，PE路由器使用该协议交换存储于VRF表中的路由信息。

BGP和它的扩展的一个特征是：它接收它的路由广告，但它不是把它们传递给每个直接相连的路由器（如IGP的例子），而是仅传递给在它的配置参数中被指定为邻居的那些路由器。注意，距离很多跳的路由器也是有可能被指定为邻居的。配置PE路由器，以便它们把所有来自客户站点的路由广告发送给那些被认为是邻居的PE路由器。它们使用MP-BGP来执行这项任务。MP-BGP属性[①]的适当选择保证了PE路由器间有目的的传播。

路由广告必须发送给哪些路由器的探测问题完全取决于该提供商支持的VPN拓扑。例如，图24-17展示了这样一个网络，在该网络中PE_1路由器将路由从VPN A的站点1的VRF表传递到VPN A连接的站点2、3、4的路由器PE_2、PE_3和PE_4，接收到的路由被输入到对应站点的VRF表中。

这样，除了从直接连接到PE站点接收到的路由外，还使用MP-BGP从该VPN其他站点接收的路由补充每个VRF表。

① 这些属性在“BGP扩展共同体属性”文件中有描述，该文件当前已成为因特网草案。

24.4.5 地址空间的无关性

如果某个节点集合从不接收来自另一个节点集合的路由信息，那么这种集合中的这些点就可能是地址无关的。

VPN的边界对路由信息传播范围的限制隔离了每个VPN的地址空间，允许它在限制范围内既可以使用公共因特网地址也可以使用根据RFC1819预留的私有地址。

在这种情况下，为什么不能在VPN内任意选择地址并且仅受TCP/IP栈采用的共同选址规则的限制？在大多数情况下，客户不希望他们的VPN完全隔离。他们中的大多数至少需要因特网访问。如果选址没有与因特网规则协调，那么内部地址可能与因特网上某个已经被使用的公共地址相匹配。从而，不可能进行因特网访问。当使用预留的私有地址时，VPN客户和外界的通信问题可以使用标准的NAT技术解决。无论如何，都必须遵守同一VPN内地址唯一性的要求。

在不同VPN内使用同一地址空间给PE路由器带来了一个新问题。BGP最初是在这样一个假设下开发的：它操作的所有地址都属于IPv4地址簇，且这些地址在整个互联网上是独一无二的。地址之所以是面向全局独一无二的是因为，收到下一个路由广告后，BGP分析它时不考虑该路由属于哪个VPN。如果对不同VPN节点的描述到达BGP输入端但包含相匹配的IPv4地址，则BGP认为它们是要前往同一节点。根据它的操作算法，BGP仅把最优路由（根据BGP选择规则选择的）写入它的VRF表。

在MPLS VPN中，该问题是使用一种新类型的扩展地址VPN-IPv4来解决的，而不是使用二义性的IPv4地址。这些扩展地址是通过转换原IPv4地址得来的。对于这个转换，为了唯一地定义该VPN的地址空间，所有的IPv4地址都添加了一个**路由标识（route distinguisher，RD）**前缀。结果，在PE路由器上所有属于不同VPN的地址都是不同的，即使它们有一个匹配部分——IPv4地址。

这是MP-BGP携带不同类型地址（包括IPv6、IPX和最重要的VPN-IPv4）的能力被证实特别有用的所在。VPN-IPv4仅用于与PE路由器使用BGP进行交换的路由。在发送明确的路由给对方前，输入PE路由器给它的IPv4目的地址增加一个RD前缀，这样就把它转换为VPN-IPv4路由了。

如前所述，RD必须唯一地标识VPN以避免地址重复。为了在不创建额外的中心程序的条件下简化RD选择（例如通过因特网权威机构来分配RD，就像IPv4地址的分配一样），假设为RD使用了保证为唯一的数值。这些数值可能是自治系统的编号或位于提供商主干上的PE接口的全局地址。

RD长为8字节，它由三个字段组成。

- *类型*——这个2字节的字段定义了类型和第二个字段的长度。
- *管理者*——该字段唯一地标识了提供商。如果*类型*字段设为0，则*管理者*字段指定PE路由器接口的IP地址；相应地，该字段的长度为4字节。如果*类型*值为1，则自治系统被选择为提供商的标识符。如果是这种情况，则*管理者*字段长为2字节。
- *编号*——该字段的目的是保证在提供商网络范围内VPN地址的唯一性。*编号*字段的值由提供商选择。它们可能是任意数值；唯一必须遵守的要求是，这些数值和提供商VPN之间的无二义性的映射。

图24-19阐释了在MPLS VPN中交换路由广告的过程。该过程包括从IPv4到VPN-IPv4的地址格式转换、路由广告过滤（导入和导出操作）和给广告添加VPN标记。

图24-19展示了一个为了保证在同一服务提供商的所有VPN框架内地址唯一性的IPv4地址转换示例。当为每个VPN创建RD时，网络管理员首先选择输入路由器PE_1的某个外部接口的全局地址（在这个例子中该地址是123.45.67.89）。为了获得RD值，管理员已分配的选择编号的值，在本例中

为1，然后使用一个冒号作为分隔符把它添加到全局地址上。这样，最终RD值为123.45.67.89:1。表24-2描述了最终的RD格式。

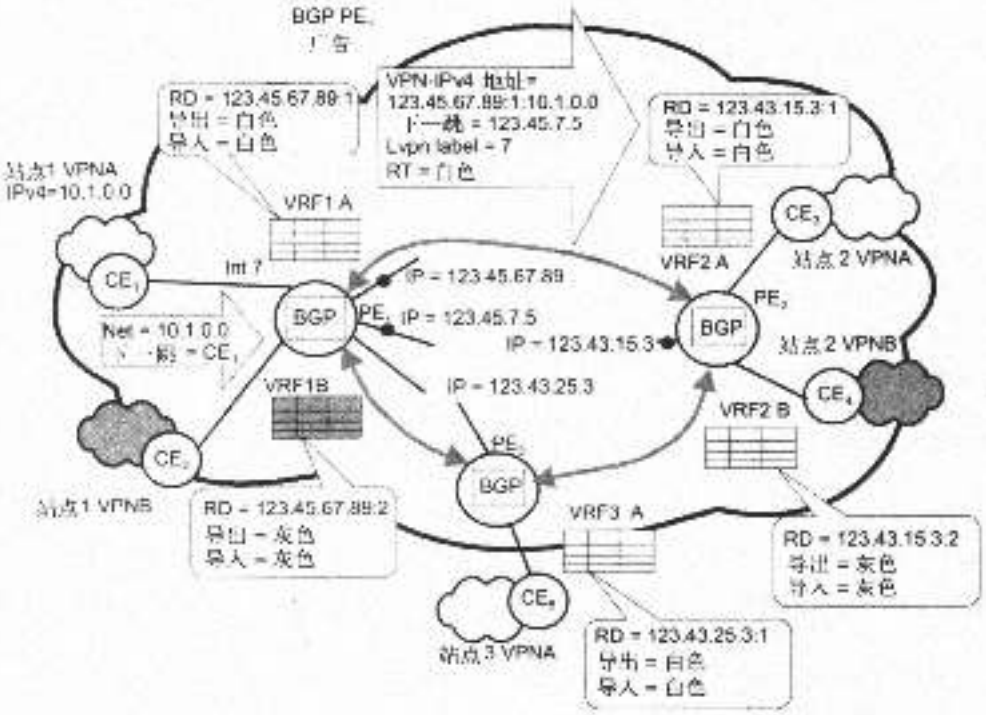


图24-19 MP-BGP路由广告

该RD值被赋给VPN A网络。当配置PE路由器时，管理员为所有对应VPN A的VRF指定该RD。实际上，当创建VRF 1A时它就指定了该值，所以对于MB-BGP协议在VRF1A表中提供的所有IPv4格式的地址的RD都设为123.45.67.89:1，包括PE_1从网络VPN A内的站点1的路由器CE_1处接收到的所有前缀为10.1/16的地址。

表24-2 RD格式

类型字段	管理者字段	编号字段
(2字节)	(4字节)	(2字节)
0	123.45.67.89	1

类似地，对于VPN B网络，管理员选择RD值123.45.67.89:2，该值在路由器PE_1上配置VRF1B时指定。当MP-BGP协议处理所有存储于VRF1B中的IPv4地址时，这个RD值将被添加到这些地址中。

说明 VRF表中的所有路由都包含IPv4格式的地址。

路由器PE1使用MP-BGP协议传递已转换为VPN-IPv4格式的路由给路由器PE_2，PE_2与VPN B网络的站点2相连。运行于远程PE路由器上的BGP协议可以区分对应于不同VPN但IPv4地址相同的路由，这要归功于附加的RD。

RFC 2547bis文件不要求同一VPN内的所有路由都由有一个RD值索引。此外，与同一PE或不同PE的不同接口相连的同一站点可以有不同的RD。因此，到同一节点的路径可以由不同的路由来描述，这提供了为不同类型分组选择某个明确的路由的可能性。然而，确保不同VPN的RD不匹配是很重要的。

24.4.6 MP-BGP路由广告的生成

当使用IGP类协议（RIP、OSPF或IS-IS）从客户站点接收到一个新路由后，PE路由器把它输入到对应的VRF表中并把它传播给该VPN的其他站点。路由信息在MP-BGP的控制下在每个单独的VPN的站点间进行交换。MP-BGP路由广告有下列属性集合，与BGP相比有所扩充：

- **VPN-IPv4格式的目的网络地址**（Address of the destination in the VPN-IPv4 format）
- **下一跳地址**（Address of the next router）（BGP下一跳）。在这种情况下运行它的BGP指定PE的一个内部（连接到P路由器）接口。
- **VPN标记**（The VPN label，LVPN）唯一地标识PE路由器的外部接口和所连接的公告路由前往的客户站点。它由输入PE赋给该路由，并被赋给一个来自连接的CE的本地路由。
- **扩展共同体属性**（Extended community attributes），它们中的**路由——目标**（route-target，RT）是强制的。这个属性标识了该VPN的这样一部分站点的集合（VRF），PE必须把路由发送给这些站点。

在路由广告中的RT属性值由配置包含当前路由的VRF表时指定的导出目标策略定义。

例如，假设PE_1路由器根据IGP-类协议从VPN A站点1接收到一条IPv4格式的路由广告（见图24-19）：

```
Net=10.1.0.0
Next-Hop=CE1
```

根据这条广告，一条对应记录被写入VRF1A表中。BGP周期性地检查VRF 1A，在发现一条新记录后，就产生一条新广告。为了实现这个功能，它执行下列操作：

- 给目的网络地址增加一个RD值（在这个例子中为123.45.67.89：1）。
- 覆盖下一跳的域的值。为了做到这点，必须进行目的地址（在本例中为123.45.7.5）经过的PE_1外部接口的地址取代CE_1的外部接口地址。
- 把一个指向VRF1A和PE_1路由器接口的标记L_{VPN}赋给包含目的节点的客户站点（在这个例子中，标记值为7，接口指定为int7）。
- 指定RT属性（在图24-19中，RT属性值按照惯例指定为WHITE，它标识了这个属于VPN A站点的结合）。

最终路由广告如下所示：

```
VPN-IPv4：123.45.67.89：1：10.1.0.0
Next-Hop=123.45.7.5
L_VPN=7
RT=White
```

MP-BGP把这条广告发送给它的所有的邻居（在图24-19中，它由一个宽箭头表示）。

当输出PE接收到到VPN-IPv4网络的路径时，它执行一个相反的转换，丢弃RD前缀，然后把它写入VRF表，接着对这个VPN的CE路由器声明该路由。这样，该VRF表中的所有路由地址格式都为*IPv4*。

表 24-3

Net	Next-Hop	L_{VPN}
10.1/16	123.45.7.5（BGP）	7

最后，在表24-3中插入下列新记录：

24.4.7 在MPLS VPN上的分组转发

既然我们已经描述过路由信息在MPLS VPN上的传播方法，现在来思考一下同一VPN的节点间数据是如何转发的。

例如，假设VPN A站点2的一个地址为10.2.1.1/16的节点发送一个分组给同一VPN的站点1的地

址为10.1.0.3/16的节点（图24-20）。使用标准传输，这个分组被传送给路由器CE_3。在那个路由器上的路由表指定路由器PE_3作为网络10.1.0.0的下一个路由器。分组由接口2传送给路由器PE_3。所以，为了给分组转发选择路径，需要到表VRF2A中查找有关该接口的地址。

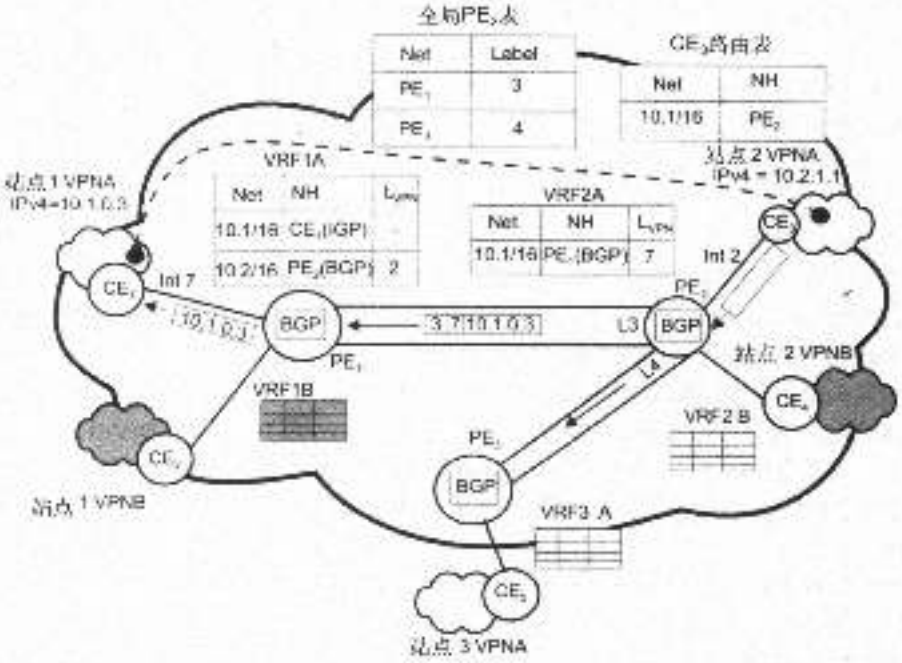


图24-20 在MPLS VPN上的分组转发

表VRF2包含了映射地址10.1.0.0到BGP记录的记录，根据这条BGP记录，路由器PE_1被指定为该分组的下一个路由器。该记录的下一个字段包含$L_{vpn}=7$，这决定了传送分组到要求的VPN路由器PE_1的接口。这条记录也说明了它是由BGP而不是IGP插入的。在此基础上，PE_3得出如下结论：下一个路由器不是直接邻居且到它的路径应在全局路由表中。

全局路由表为PE_3能够指定LSP路径的L标记的初始值。在本例中，这个值为3。我们将不再把注意力放在PE_1和PE_2间创建路径的方法上，因为这个话题已在第22章讨论过了。

MPLS VPN技术使用MPLS路径的层次特性。使用这些特性可以提供分组多个位于栈内的标记。当进入由P路由器组成的提供商内网时，将提供分组两个标记：内部标记$L_{vpn}=7$和外部标记$L=3$。标记L_{vpn}解释为低层标记。当分组穿过PE_1-PE_2隧道时，它留在栈底不被使用。分组转发基于高层标记，L标记。每次分组经过隧道的下一个P路由器时，都对L标记进行分析并用一个新值替换。只有在分组到达隧道终点PE_1路由器后，才从栈中获取L_{vpn}标记。根据它的值，分组被转发给要求的PE_1路由器的输出接口int7。

然后，从VRF1A表中关于该接口的记录中获得指定CE_1为下一个路由器的到目的节点路由记录，且该记录包含VPN A路由。注意到这条记录是由IGP插入到VRF1A表中的。在分组由CE_1到节点10.1.0.3传送的最后部分，是使用传统IP方法转发的。

24.4.8 形成VPN拓扑的机制

导出/导入策略是创建不同VPN拓扑的有力工具。

当配置VRT表时，要指定两个RT属性，一个用于定义导出策略，另一个用于定义导入策略。

MP-BGP路由广告通常要带指定路由导出策略的RT属性。通过比较路由广告和VRF参数中的

RT属性值，可以决定该路由是必须接受还是必须拒绝。这个方法是形成网络拓扑的方法。用一个具体示例来说明它。

假设路由器PE_2（见图24-19）从PE_1处收到一条广告。在保存有关该路由的信息前，它检查广告中的RT属性是否与它的任何一个VRF表的导入策略匹配（在本例中为VRF2A和VRF2B）。RT属性值为WHITE；因为WHITE导出策略仅由表VRF2A定义，所以，在转换为IPv4格式（丢弃RD）后，这个路由仅被插入到表VRF2A中。表VRF2B保持不变，因为它的策略说明只有RT属性为GRAY的路由才必须输入。

为一具体VPN的所有VRF的导出和导入策略指定同一个值（如图24-19中的VPN A）会产生一个全连通。每个站点把它的分组直接发送给属于该目的网络的站点。

VPN拓扑还有其他变体。例如，通过配置导出/导入策略，可以实现“星型”（中心辐射，hub and spoke）这样的流行拓扑。在这个拓扑中，所有的站点（辐射，spoke）通过一个专门的中心站点（中心，hub）进行通信。

为了实现这个效果，为中心站点的VRF定义策略为：输入=辐射（import=spoke）即可。导出策略必须定义为输出=中心（export=hub）。至于外围站点的VRF，则必须转换策略，指定输入=中心（import=hub），输出=辐射（export=spoke）（图24-21）。结果，外围站点的VRF将不接收彼此间的路由广告，因为这些路由目标属性设为辐射的广告是使用MP-BGP在网络上传送的，且它们的导入策略仅允许它们接收RT=中心的路由广告。另一方面，来自外围站点的VRF广告可以被中心站点的VRF接收，因为它们的导入策略被设为辐射。中心站点归纳所有来自外围设备的广告并把它们发送回去；然而，这次RT属性被设为匹配外围站点导入策略的中心。因而，每个外围站点的VRF都补充以网络上其他外围站点的带有连接到指定为下一跳的中心站点的PE接口地址的记录（因为这个接口是接收广告的接口）。所以，外围站点间的分组将经过连接到中心站点的路由器PE_3。

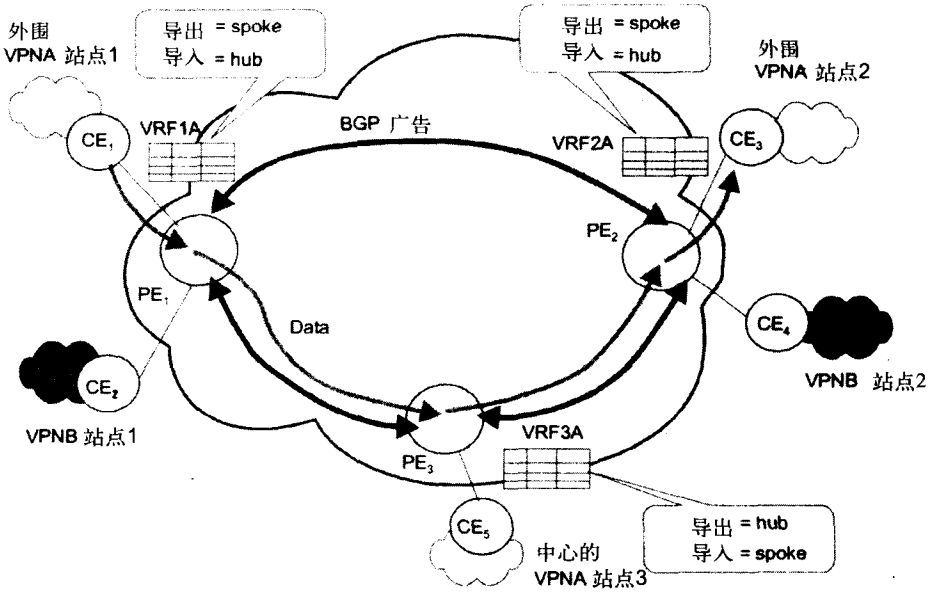


图24-21　为VPN A配置星型拓扑

基于对MPLS VPN方法的描述，可以说配置一个新VPN或修改已有VPN配置的程序是复杂的。然而，该程序可以形式化或自动化。为了减少可能的配置错误——如给站点赋予了错误的导入/导出策略，这会连接到外部VPN站点——一些制造商已为MPLS配置开发了自动化软件。

24.4.9 安全水平

MPLS VPN的安全水平可以使用传统的方法进行改进——例如，使用安装于客户或提供商网络上的认证和加密IPSec工具。MPLS VPN服务可以轻易地与其他IP服务集成，如提供因特网访问VPN用户并使用安装于提供商网络的防火墙保护它们。 提供商可以根据其他MPLS能力给MPLS VPN用户一些服务——如，在MPLS流量工程方法的基础上确保有保障的QoS。 至于支持在提供商路由器上的用户路由表的困难（为有些专家所强调的），依我们所见是有点言过其实的。毕竟，路由表是使用标准路由协议自动创建的。此外，它们仅创建于PE路由器上。虚拟路由方法把这些表和提供商全局路由表隔离起来，这保证了所需的MPLS VPN的可延拓性和可靠性。尽管如此，未来这项技术还将展示它的真正特性——这可能在不久的将来就会实现。

MPLS VPN不能像在IPSec和PPP中那样使用加密和认证来保证安全。然而，需要时它允许使用这些技术作为附加的保护方法。

MPLS VPN并没有将保证QoS支持作为自己的目标；然而，当需要时，提供商也可以使用DiffServ和MPLS流量工程方法。

小结

- 安全传输服务允许客户在公共网中传输流量并能保证所传信息真实性、完整性和机密性的安全方法，如使用因特网传送。
- 安全服务可以建立在实现于OSI模型各层上的系统工具的基础上。最流行的安全传输服务有SSL，IPSec，PPTP和VPN。
- IPSec是开放标准的一个协商集，如今它有包括三个协议的严格的核心：保证数据完整性和真实性的AH，保证数据完整性和认证并使用加密保证数据机密性的ESP，建立逻辑连接并分配私钥的IKE。
- 当建立被称为IPSec安全关联（SA）的单向逻辑连接时，参与方对决定数据传输的多个参数进行协商并达成一致：类型和安全协议的运行模式（AH或ESP）、加密方法和私钥等等。决定所有活动关联的当前参数集合以SAD格式存储于安全信道的两个端节点。每个参与方还支持另一个数据库SAD，它指定IP分组属性和为它们建立的处理规则之间的映射。
- 为了流量保护而使用SPD和SAD，考虑了建立带有IP流量数据报性质的逻辑连接的SA方法的灵活组合。
- VPN技术允许多个公司间共享同一网络基础设施的实现接近于使用有保证的QoS（安全性、可用性、可预测带宽和独立寻址）的真实专用网提供的那些服务。
- VPN可以由某家公司（CPVPN）或某个服务提供商（PPVPN）实现。它可以基于用户驻地设备也可以是基于安装于提供商领域的设备。
- VPN技术可以根据用于保证数据传输安全的方法分为两类：流量分离技术（ATM VPN、帧中继VPN和MPLS VPN）和流量加密技术（IPSec VPN）。
- MPLS VPN较之其他VPN，如建于ATM、帧中继或IPSec基础上的VPN的优点，是它的高可延拓性，自动配置的可能性和与其他IP服务的自然结合。
- 为属于同一VPN的站点所使用的、用于交换路由信息的方法是多协议BGP（MP-BGP）扩充。
- 路由广告的导出/导入策略是创建不同拓扑MPLS VPN的一个有力工具。

复习题

1. 可以为不引入任何变化的应用产生的流量的安全数据传输使用IPSec吗？
2. IPSec系统为保证数据完整性和认证提供了AH协议。那么它又是出于什么目的提供了ESP协议，又是谁实现这些功能呢？

3. 为了提供接收端检查数据完整性的能力，大多数协议在分组中设置有校验和。IPSec为了保证数据完整性使用了摘要。解释这两种方法的区别。
4. 假设在你的计算机上运行着三个应用程序，且你需要使用IPSec以加密形式传送这三个程序产生的数据给你的搭档。为了实现这个目的，你需要创建多少个SA?
5. 比较IPSec的传输和隧道模式。哪一个保证了更高的安全水平？哪一个提供更好的可延拓性？哪一个更经济?
6. 列举多个例子来说明入侵者使用头部信息的方法。
7. 与在传输模式下使用AH协议相比，在隧道模式下使用这个协议能否提高所传数据的安全性?
8. 在IPSec使用的是哪种保护方法来防止入侵者的复制?
9. 解释为什么说填充是保证机密性的另一种方法。
10. SG是如何判断抵达分组所需的处理类型?
11. VPN可以支持真实专用网的那些特性?
12. 提出一种VPN分类。
13. 哪种VPN技术使用流量分离保证安全性?
14. L3VPN与L2VPN相比有哪些优缺点?
15. IPSec VPN的主要缺点是什么?
16. 叙述MPLS VPN中不同的客户网络地址空间的分离方法?
17. VRF表是如何创建的?
18. 假设运行于路由器PE2上MP-BGP发送了一条路由广告，该路由广告是由从路由器CE_3接收到的广告转换来的（图24-20）：

    ```
    Net=10.2/16
    Next-Hop=CE3
    ```

    这条路由广告是如何产生的?
19. 在MPLS VPN网络中，分组被提供两个标记，内部标记（L_{VPN}）和外部标记（L）。描述这些标记在分组转发中的作用?

练习题

1. 你已经学习了当所有客户的所有站点都连接到同一个提供商的主干时，L3MPLS VPN的运行原理。当主干是由多个提供商支持时，尝试改进这些原理。假设三个客户A、B和C的站点是连接到提供商ISP2和ISP3的网络上的。而这些提供商的网络是使用ISP1提供商网络连接的。使用一个分层的方法来解决这个问题，把ISP1作为最顶层提供商。在这种情况下，ISP2和ISP3将根据本章描述的MPLS VPN方法扮演ISP1的客户的角色。以BGP的能力和MPLS标记栈的思想作为基础，给出一种MPLS VPN提供商层次思想的可能的实现。
2. 比较下列两种情况所需的虚电路的数目和VPN服务提供商必须创建的LSP的数目：
   - 提供商使用帧中继网络提供VPN服务
   - 提供商使用IP/MPLS网络提供VPN服务

   假设提供商有25个客户，每个客户的网络包含10个连接到提供商网络的站点。客户需要企业内部网服务，哪种方法不需要提供客户站点间的连接?
3. 在24.4.7节中，有一个说明从VPN A站点2的节点10.2.1.1/16到位于同一个VPN内的站点1的端节点10.1.0.3/16的分组传输的图例（图24-20）。使用这个插图，描述相反方向，即从节点10.1.0.3 /16到节点10.2.1.1/16的分组传输。提供CE_1和PE_1路由表中的可能的内容。对默认数据给出你的值，并把它们写入到图例中。

结束语　展望未来

我们把目光放得越远，我们看到传统意义下的计算机网络——仅仅传送文本和数字的网络——的机会越小。无论是电话网、计算机网、还是电缆电视网，各种网络发展的趋势都是融合。现在计算机网络已经开始传送各种各样的信息流，而不仅是最初计算机网络典型的信息流。这会以各种各样的形式出现，比如，以电话网中两个用户间的传统的交互式通话的形式、以通过因特网按需广播（音乐、预先录制的讲话或面谈）的形式、以语音邮件的形式等。图像的传输被认为要求较高的网络吞吐能力因而相对较少。但是，甚至在约64～128 Kb/s的访问速度下，在PC屏幕上一个小矩形窗口中观看电视广播已成为可能。

因此，未来的电信网络不但能够传送突发的数据流，也同样能够传送声频和视频流。未来的网络将继承其前任（电话、计算机、无线电和TV广播网等）最好的特性。但是，它们将使用共同的运输技术，并确保以要求的QoS传输各种流量。根据大多数专家的共同观点，这种技术的实现必须基于分组交换技术并广泛采用取得胜利的协议——IP。这使得将来的网络类似于现在的计算机网络；当然，尚期待若干重要的技术创新。

这些技术创新可能包括新型的终端设备，这些设备需既具备PC的功能，又具备电话设备易于使用和简单的特点。智能电话和PDA是这类设备的原型。这样的设备让使用者通过按几个按钮就可以访问预定的网页、进行电话交谈、发送附带多媒体应用的电子邮件消息、请求视频点播（以及其他更多目前还仅在研究项目中存在的服务）。这些新设备的出现将给电信业的革新带来动力。

基于DWDM和通用MPLS（GMPLS）标准的控制虚通路技术将可以应对超高速和高质量运输要求的增长。新的公众电信网络的核心将基于多核光纤电缆。它能够保证通信节点间几兆兆字节（terabytes）的吞吐量，也将构成在今天看来似乎难以想像的信息量传输的基础。从经济角度考虑，这个核心必须支持只是超高速数据流的交换，如仅是某特定波长（DWDM交换）甚或是单核心的数据流——没有较小的交换单位。结果，SDH技术将从该网络核心中退出，而扮演DWDM交换网的接入网的角色。另一个革新性的成就将是，当成员核心的通路、波长和SDH容器使用统一信令协议动态生成时，基于GMPLS技术的网络核心可控性。最重要的是这种协议将有一个终端用户版本。这意味着核心用户（比如，服务提供者）能够根据当前的需要灵活使用吞吐量。

当前，较低的访问速率，是阻碍新的多媒体服务推广的主要障碍之一，特别对广泛的用户团体而言。对此有多种解决方案，包括采用已有的铜质本地回路（对大多数个人用户最适合的方式）、采用固定或移动的无线访问接入以及使用经济的无源光纤网络技术安装光纤本地回路。ATM或IP/MPLS技术将被用来分享信道带宽。

尽管网络核心和接入网的吞吐率都有了相当的增长，但当同时发生的流量超出了网络连接的能力时，通信拥塞仍有可能发生。因此，为实现高质量的流量传输，将来的网络将广泛采用QoS支持方法。在网络核心部分，这些将是确保为大量用户携带数据的巨大汇聚数据流服务的方法，换句话说，这些方法类似于在承载网中开始找到应用的DiffServ。在接入网部分，类似于ATM和IntServ的方法将为个体流提供服务。

LAN也在变化。连接计算机的无源电缆正被各种通信设备（如交换机、路由器和网关）所取代。由于这些设备的使用，通过复杂的结构连接构建数以千计计算机的大型协同网络变成可能。随着由PC易于操作的特点带来的惊喜逐渐减弱，主要由于成千上万台服务器组成的系统较几台大型机更难以维护，对大型超级计算机的兴趣又逐渐复苏。因此，主机在新一轮的技术革新中又重新回到企业中。但是这一次，它们成为网络的完全成员，支持以太网或令牌环技术和TCP/IP栈，而TCP/IP栈由于因特网成为了事实上的标准。

这些仅是电信网络发展的一些方向，但有的在今天已经可以清楚看到。

参考文献与推荐阅读的书

第一部分推荐阅读的书

1. Armitage, G. *Quality of Service in IP Networks.* Pearson Education, 2000.
2. Black, U. *Internet Security Protocols: Protecting IP Traffic,* Ed.1. Prentice Hall, 2000.
3. Black, U. *Emerging Communications Technologies,* Ed. 2. Prentice Hall Professional, 1997.
4. Black, U. *Data Networks: Concepts, Theory and Practice.* Englewood Cliffs, New Jersey: Prentice Hall, 1989.
5. Comer, D.E. *Internetworking with TCP/IP,* Vol. 1: *Principles, Protocols, and Architecture,* Ed. 3. Prentice Hall, 2000.
6. Comer, D.E., Stevens, D. *Internetworking with TCP/IP,* Vol. 2: *Design Implementation, and Internals.* Prentice Hall, 1994.
7. Comer, D.E., Stevens, D. *Client-Server Programming and Applications.* Prentice Hall, 2001.
8. Dodd, A.Z. *The Essential Guide to Telecommunications 1,* Ed. 2. Prentice Hall, 1999.
9. Dorogovtsev, S.N., Ferreira Mendes, J.F. *Evolution of Networks: From Biological Nets to the Internet and WWW.* Oxford University Press, 2003.
10. Fitzgerald, J. *Business Data Communications.* John Wiley & Sons, 1993.
11. Ford, W. *Computer Communications Security.* Prentice Hall, 1994.
12. Freeman, R. *Telecommunications Transmission Handbook.* NY: Wiley, 1998.
13. Hamacher, C.V., Vranesic, Z.G., Zaky, S.G. *Computer Organization.* New York: McGraw Hill, 1984.
14. Halsall, F. *Data Communications, Computer Networks, and Open Systems.* Addison-Wesley, 1996.
15. Hardy, W.C. *QoS Measurement and Evaluation of Telecommunications Quality of Service.* John Wiley & Sons, 2001.
16. Hauben, M., Hauben, R., Truscott, T. *Netizens: On the History and Impact of Usenet and the Internet,* Ed. 1. Wiley–IEEE Computer Society Pr., 1997.
17. Ibe, O.C. *Converged Network Architectures: Delivering Voice and Data Over IP, ATM, and Frame Relay.* Wiley, 2001.
18. Keshav, S. *Efficient Implementation of Fair Queuing.* Proceedings of the ACM SIGCOMM, 1990.
19. Keshav, S. *An Engineering Approach to Computer Networking.* Addison Wesley, 1997.
20. Kleinrock, L. *Queuing Systems,* Vol. 1: *Theory.* Wiley-Interscience, 1975.
21. Kurose, J.F., Ross, K.W. *Computer Networking: A Top-Down Approach Featuring the Internet,* Ed. 3. Addison Wesley, 2004.
22. LaQuey, T. *The Internet Companion: A Beginner's Guide to Global Networking.* Reading, Massachusetts: Addison-Wesley, 1994.
23. Leiner, B.M., Cerf, V.G., et al. *A Brief History of the Internet.* Internet Society (ISOC), **www.isoc.org/internet/history/brief.shtml**.
24. Leon-Garcia A., Widjaja, I. *Communication Networks: Fundamental Concepts and*

Key Architectures, Ed.1. McGraw Hill Science/Engineering/Math, 2001.

25. Le Boudec, J.-Y., Thiran, P. *Network Calculus. A Theory of Deterministic Queuing Systems for the Internet Series: Lecture Notes.* In: *Computer Science*, Vol. 2050, 2001.
26. Null, L., Lobur, J. *The Essentials of Computer Organization and Architecture Dimensions.* Jones & Bartlett Pub, 2003.
27. Osborne, E., Ajay, S. *Traffic Engineering with MPLS.* Cisco Press, 2002
28. Peterson, L.L., Davie, B.S. *Computer Networks: A Systems Approach*, Ed. 3. Morgan Kaufmann, 2003.
29. Rose, M.T. *The Open Book: A Practical Perspective on OSI.* Englewood Cliffs, New Jersey: Prentice Hall, 1990.
30. Rowe, S.H. *Telecommunications for Managers*, Ed. 3. Prentice Hall, 1995.
31. Stallings, W. *Data and Computer Communications*, Ed. 7. Prentice Hall, 2003.
32. Stallings, W. *Wireless Communications and Networks.* Prentice Hall, 2002.
33. Stallings, W. *Local and Metropolitan Area Networks.* Macmillian Publishing Company, 1993.
34. Standards. International Organization for Standardization. Information Processing System — *Open System Interconnection: Specification of Abstract Syntax Notation One (ASN.1).* International Standard 8824, 1987.
35. Stevens, R.W. *TCP/IP Illustrated, Vol. 1. The Protocols*, Ed. 1. Addison-Wesley Professional, 1993.
36. Young Moo Kang, Miller, B.R., Pick, R.A. *Comments on "Grosch's law re-revisited: CPU power and the cost of computation".* Communications of the ACM, Vol. 29, Issue 8, 1986.
37. Zheng Wang. *Internet QoS: Architectures and Mechanisms for Quality of Service*, Ed. 1. Morgan Kaufmann, 2001.
38. Zwcky, E.D., Cooper, S., Chapman, B.D. *Building Internet Firewalls.* O'Reilly, 2000.

第二部分推荐阅读的书

1. Abbas, J. *The Wireless Mobile Internet.* John Wiley & Sons, 2003.
2. Ashwin, G. *DWDM Network Designs and Engineering Solutions.* Pearson Education, 2002.
3. Bertsekas D., Gallager, R. *Data Networks*, Ed. 2. Prentice Hall, 1992.
4. Black, U. *Physical Level Interfaces and Protocols.* Los Alamitos, California: IEEE Computer Society Press, 1988.
5. Couch, L.W. *Digital and Analog Communication Systems*, Ed. 6. Prentice Hall, 2001.
6. EIA232E p: *Interface Between Data Terminal Equipment and Data Circuit-Terminating Equipment Employing Serial Binary Data Interchange*, revised from EIA232D, July 1991.
7. Halsall, F. *Data Communications, Computer Networks, and Open Systems.* Addison Wesley, 1996.
8. Gibson, J.D. *Principles of Digital and Analog Communications*, Ed. 2. Prentice Hall, 1993.
9. Haykin, S. *Digital Communications.* Wiley, 1988.
10. Nicopolitidis, P., Obaidat, M.S., et al. *Wireless Network.* Wiley, 2003.
11. Proakis, J.G. *Digital Communications*, Ed. 4. McGraw Hill, 2001.
12. Rowe, S.H. *Telecommunications for Managers*, Ed. 3. Prentice Hall, 1995.

13. Sexton, M., Reid, A. *Broadband Networking: ATM, SDH, and SONET.* Artech House, 1997.
14. Shu Lin, Costello, D.J. *Error Control Coding,* Ed. 2. Prentice Hall, 2004.
15. Sklar, B. *Digital Communications: Fundamentals and Applications,* Ed. 2. Prentice Hall PTR, 2001.
16. Stallings, W. *Data and Computer Communications,* Ed. 6. Prentice Hall, 2004.
17. Stallings, W. *Wireless Communications and Networks.* Prentice Hall, 2002.
18. Sweeney, P. *Error Control Coding.* Wiley, 2002.
19. Wicker, S.B. *Error control systems for digital communication and storage.* Prentice Hall, 1995.
20. Wirth, N. *Digital Circuit Design for Computer Science Students. An Introductory textbook.* Springer Verlag, 1995.
21. Zeimer, R.E., Peterson, R.L. *Introduction to Digital Communication.* Prentice Hall, 2001.

第三部分推荐阅读的书

1. Abbas, J. *The Wireless Mobile Internet.* John Wiley & Sons, 2003.
2. Bertsekas D., Gallager, R. *Data Networks,* Ed. 2. Prentice Hall, 1992.
3. Black, U. *Data Networks: Concepts, Theory and Practice.* Englewood Cliffs, New Jersey: Prentice Hall, 1989.
4. Bux, W. *Local-area subnetworks: a performance comparison.* IEEE Press Trans. Comm., vol. COM-29, Issue 10, 1981.
5. Cunningham,D., Lane, W.G., Lane, B. *Gigabit Ethernet Networking,* Ed. 1. Sams, 1999.
6. Gast M., Gast, M.S. *802.11 Wireless Networks: The Definitive Guide,* Ed. 1. O'Reilly, 2002.
7. Halsall, F. *Data Communications, Computer Networks, and Open Systems.* Addison Wesley, 1996.
8. Hillston, J.E., King, P.J.B., Pooley, R.J., editors. *Computer and Telecommunications Performance Engineering.* London: Springer Verlag, 1992.
9. Metcalfe, R.M., Boggs, D.R. *Ethernet: Distributed packet switching for local computer networks.* Comm. ACM 19, 7, 1976.
10. Miller, B.A., Bisdikian, C. *Bluetooth Revealed: The Insider's Guide to an Open Specification for Global Wireless Communications,* Ed. 2. Prentice Hall, 2001.
11. McNamara, J.E. *Local Area Networks.* Bedford, MA: Digital Press, Educational Services, 1985.
12. Norris, M. *Gigabit Ethernet Technology and Applications.* Artech House Publishers, 2002.
13. Perlman, R. *Interconnections: Bridges, Routers, Switches, and Internetworking Protocols,* Ed. 2. Addison-Wesley Professional, 1999.
14. Peterson, L.L, Davie, B.S. *Computer Networks: A Systems Approach,* Ed. 3. Morgan Kaufmann, 2003.
15. Riley, S., Breyer, R. *Switched, Fast, and Gigabit Ethernet,* Ed. 3. Sams, 1998.
16. Ross, F.E. *FDDI—A Tutorial.* IEEE Communications Magazine, Vol. 24, No. 5, 1986.
17. Schwartz, M. *Telecommunications Networks — Protocols, Modeling and Analysis,* Facsimile Ed. Addison Wesley, 1986.
18. Seifert, R. *Gigabit Ethernet: Technology and Applications for High-Speed LANs,* Ed. 1.

Addison-Wesley Professional, 1998.

19. Seifert, R. *The Switch Book: The Complete Guide to LAN Switching Technology*, Ed. 1. Wiley, 2000.
20. Spurgeon, C.E. *Ethernet: The Definitive Guide.* O'Reilly, 2000.
21. Stallings, W. *Data and Computer Communications*, Ed. 6. Prentice Hall, 2004.
22. Stallings, W. *Wireless Communications and Networks.* Prentice Hall, 2002.

第四部分参考文献

[RFC 751] Lebling, P. *Survey of FTP mail and MLFL.* RFC 751, 1978.

[RFC 760] Postel, J. *DoD standard Internet Protocol.* RFC 760, 1980.

[RFC 768] Postel, J. *User Datagram Protocol.* 1980.

[RFC 791] Postel, J. *Internet Protocol.* STD 5, RFC 791, 1981.

[RFC 792] Postel, J. *Internet Control Message Protocol.* 1981.

[RFC 793] Postel, J. *Transmission Control Protocol.* 1981.

[RFC 950] Mogul, J., Postel, J. *Internet Standard Subnetting Procedure*, STD 5, RFC 950, 1985.

[RFC 1122] Braden, E.R. *Requirements for Internet Hosts — Communication Layers.* 1989.

[RFC 1349] Almquist, P. *Type of Service in the Internet Protocol Suite.* 1992.

[RFC 1517] Internet Engineering Steering Group and Hinden, R. *Applicability Statement for the Implementation of Classless Inter-Domain Routing (CIDR).* RFC 1517, 1993.

[RFC 1518] Rekhter, Y., Li, T. *An Architecture for IP Address Allocation with CIDR.* RFC 1518, 1993.

[RFC 1519] Fuller, V., Li, T., Yu, J., Varadhan, K. *Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy.* RFC 1519, 1993.

[RFC 1520] Rekhter, Y. Topolcic, C. *Exchanging Routing Information Across Provider Boundaries in the CIDR Environment.* RFC 1520, 1993.

[RFC 1700] Reynolds J., Postel, J. *Assigned Numbers.* 1994.

[RFC 1752] Bradner, S., Mankin, A. *The Recommendation for the IP Next Generation Protocol.* RFC 1752, 1995.

[RFC 1878] Pummill, T., Manning, B. *Variable Length Subnet Table For IPv4.* RFC 1878, 1995.

[RFC 2050] Hubbard, K., Kosters, M., et al. *Internet Registry IP Allocation Guidelines.* BCP 12, RFC 2050, 1996.

[RFC 2131] Droms, R. *Dynamic Host Configuration Protocol.* RFC 2131, 1997.

[RFC 2132] Alexander, S., Droms, R. *DHCP Options and BOOTP Vendor Extensions.* RFC 2132, 1997.

[RFC 2373] Hinden, R. *IP Version 6 Addressing Architecture.* RFC2373, 1998.

[RFC 2460] Deering, S., Hinden, R. *Internet Protocol, Version 6 (IPv6) Specification.* RFC 2460, 1998.

[RFC 2998] Bernet, Y., Ford, P., et al. *A Framework for Integrated Services Operation over Diffserv Networks.* 2000.

[RFC 3232] Reynolds, J. *Assigned Numbers: RFC 1700 is Replaced by an On-line Database*, Ed. 2002.

[RFC 3246] Davie, B., A. Charny, A., et al. *An Expedited Forwarding PHB (Per-Hop Behavior).* 2002.

[RFC 3290] Bernet, Y., Blake, S., et al. *An Informal Management Model for Diffserv Routers.* 2002.

[RFC 3513] Hinden, R., Deering, S. *Internet Protocol Version 6 (IPv6) Addressing Architecture.* RFC 3513, 2003.

第四部分推荐阅读的书

1. Boney, J. *Cisco IOS in a Nutshell.* O'Reilly, 2001.
2. Coltun, R. *OSPF: An Internet Routing Protocol.* ConneXions: The Interoperability Report, Vol. 3, No. 8, 1989.
3. Comer, D.E. *Internetworking with TCP/IP.* Vol. 1: *Principles, Protocols, and Architecture,* Ed. 3. Prentice Hall, 2000.
4. Davidson, J. *An Introduction to TCP/IP.* New York: Springer Verlag, 1992.
5. Feit, S. *Architecture, Protocols, and Implementation with IPv6 and IP Security.* McGraw Hill, 1997.
6. Hunt, C. *TCP/IP Network Administration,* Ed. 2. O'Reilly, 1998.
7. Kleinrock, L. *Communication Nets: Stochastic Message Flow and Delay.* New York: McGraw Hill, 1964.
8. Kleinrock, L. *Queueing Systems.* Vol. 2: *Computer Applications.* New York: John Wiley & Sons, 1976.
9. Kleinrock, L. Queueing Systems, Vol. 1: *Theory.* Wiley-Interscience, 1975.
10. Nogl, M. *Illustrated TCP/IP. A graphic Guide to the Protocol Suite.* John Wiley & Sons, 1999.
11. Roberts, L.G. *Judgment Call.* Data Communications magazine, April 1999.
12. Shenker, S., Partridge, C. *Specification of Guaranteed Quality of Service.* IETF draft, 1995.
13. Snader, J. *EffectiveTCP/IP programming.* DMK Press, 2001.
14. Specification: *NetWare Link Services Protocol (NLSP),* Revision 0.9. Part Number 100-001708-001. 1993.
15. Stevens, W.R. *TCP/IP Illustrated.* Vol. 1: *The Protocols,* Ed. 1. Addison-Wesley Professional, 1993.
16. Varghese, G. *Network Algorithmics: An Interdisciplinary Approach to Designing Fast Networked Devices.* Morgan Kaufmann, 2004.

第五部分参考文献

[RFC 2514] Noto, M., Spiegel, E., Tesink, K. *Definitions of Textual Conventions and OBJECT-IDENTITIES for ATM Management.* **ftp://ftp.isi.edu/in-notes/rfc2514.txt**, 1999.

[RFC 2515] Tesink, K. *Definitions of Managed Objects for ATM Management.* **ftp://ftp.isi.edu/in-notes/rfc2515.txt**, 1999.

[RFC 2684] Grossman, D., Heinanen, J. *Multiprotocol Encapsulation over ATM Adaptation Layer 5.* **ftp://ftp.isi.edu/in-notes/rfc2684.txt**, 1999.

[RFC 2761] Dunn, J., Martin, C. *Terminology for ATM Benchmarking.* **ftp://ftp.isi.edu/in-notes/rfc2761.txt**, 2000.

[RFC 2955], Rehbehn, K., Nicklass, O., Mouradian, G. *Definitions of Managed Objects for Monitoring and Controlling the Frame Relay/ATM PVC Service Interworking Function.* **ftp://ftp.isi.edu/in-notes/rfc2955.txt**, 2000.

[RFC 3035] Davie, B., Lawrence, J., et al. *MPLS using LDP and ATM VC Switching.* **ftp://ftp.isi.edu/in-notes/rfc3035.txt**, 2001.

[RFC 3116] Dunn, J., Martin, C. *Methodology for ATM Benchmarking.* **ftp://ftp.isi.edu/in-notes/rfc3116.txt**, 2001.

[RFC 3134] Dunn, J., Martin, C. *Terminology for ATM ABR Benchmarking.* **ftp://ftp.isi.edu/in-notes/rfc3134.txt**, 2001.

第五部分推荐阅读的书

1. Aboba, B. *NAT and IPSEC.* Internet Engineering Task Force, 2000.
2. *Advanced MPLS Design and Implementation.* Cisco Press, 2001.
3. Armitage, G. *Quality of Service in IP Networks.* Pearson Education, 2000.
4. Balaji, K. *Broadband Communications.* McGraw Hill, 1998.
5. Berkowitz, H. *Requirements Taxonomy for Virtual Private Networks.* Internet Engineering Task Force, 1999.
6. *Big Book of Multiprotocol Label Switching RFCs.* Morgan Kaufmann Publishers, 2000.
7. Black, U. *Internet Security Protocols: Protecting IP Traffic,* Ed.1. Prentice Hall, 2000.
8. Black, U. *Emerging Communications Technologies,* Ed. 2. Prentice Hall Professional, 1997.
9. *Building Switched Networks: Multilayer Switching, QoS, IP Multicast, Network Policy, and Service Level Agreements.* Ed. 1. Addison-Wesley, 1999.
10. Busschbach, P.B. *Toward QoS-Capable Virtual Private Networks.* Bell Labs Technical Journal, Vol. 3, No. 4, 1998.
11. Casey, L. *An extended IP VPN Architecture.* Internet Engineering Task Force, 1998.
12. Dobrowski, G., Grise, D. *ATM and Sonet Basics.* APDG Publishing, 2001.
13. Dutton, H., Lenhard, P. *Asynchronous Transfer Mode (ATM) Technical Overview,* Ed. 2. Prentice Hall, 1995.
14. Feit, S. *Architecture, Protocols, and Implementation with IPv6 and IP Security.* McGraw Hill, 1997.
15. Ford, W. *Computer Communications Security.* Prentice Hall, 1994.
16. Ginsburg, D. *ATM Solutions for Enterprise Internetworking,* Ed. 2. Addison-Wesley, 1998.
17. Gonsalves, M. *Voice Over IP Networks,* Bk & CD-Rom edition. McGraw Hill Osborne Media, 1998.
18. Hunt, C. *TCP/IP Network Administration,* Ed. 2. O'Reilly, 1998.
19. Ibe, O.C. *Converged Network Architectures: Delivering Voice and Data Over IP, ATM, and Frame Relay.* Wiley, 2001.
20. Ibe, O.C. *Essentials of ATM Networks and Services.* Addison-Wesley, 1997.
21. Jamieson, D., Jamoussi, B., et al. *MPLS VPN Architecture.* Internet Engineering Task Force, 1998.
22. Kompella, K., et al. *MPLS-based Layer 2 VPNs.* Internet Engineering Task Force, 2000.
23. Kurose, J. F., Ross, K.W. *Computer Networking: A Top-Down Approach Featuring the Internet,* Ed. 3. Addison Wesley, 2004.

24. Li, T. *CPE based VPNs using MPLS.* Internet Engineering Task Force, 1998.
25. McDysan, D.E, Spohn, D.L. *ATM Theory and Applications.* McGraw Hill, 1998.
26. McDysan, D.E, Spohn, D.L. *Hands-On ATM.* McGraw Hill, 1998.
27. Morris, S. *Network Management, MIBs and MPLS: Principles, Design and Implementation.* Prentice Hall, 2003.
28. *MPLS and VPN Architectures,* Vol. 1. Cisco Press, 2000.
29. *MPLS and VPN Architectures,* Vol. 2. Cisco Press, 2003.
30. Muthukrishnan, K., Malis, A. *Core MPLS IP VPN Architecture.* Internet Engineering Task Force, 2000.
31. Perros, H.G. *An Introduction to ATM Networks.* Wiley, January 2001.
32. Sackett, G.C., Metz, C. *ATM and Multiprotocol Networking.* McGraw Hill, 1997.
33. Siu, S., Jain, R. *A brief overview of ATM: Protocol Layers, LAN Emulation and Traffic Management.* Computer Communications Review (ACM SIGCOMM), 1995.
34. Stanford, H. *Telecommunications for Managers,* Ed. 3. Prentice Hall, 1995.
35. Stevens, R.W. *TCP/IP Illustrated,* Vols. 1, 2, 3, Ed. 1. Addison-Wesley Professional, 1993.
36. Sun, W., Bhaniramka, P., Jain, R. *Quality of Service Using Traffic Engineering over MPLS: An Analysis.* Proc. 25th Annual IEEE Conference on Local Computer Networks (LCN 2000), Tampa, Florida USA, November 8–10, 2000.
37. *The MPLS Primer: An Introduction to Multiprotocol Label Switching.* Prentice Hall, 2001.
38. Thompson, R.A. *Telephone Switching Systems,* Ed. 1. Artech House Publishers, 2000.
39. Zwcky, E.D., Cooper, S., Chapman, B.D. *Building Internet Firewalls.* O'Reilly, 2000.
40. Zorn, G., Pall, G., et al. *Point-to-Point Tunneling Protocol (PPTP).* Internet Engineering Task Force, 1999.