层次交换网络体系结构

钱华林　葛敬国　李　俊　等著

清华大学出版社

层次交换网络体系结构

钱华林　葛敬国　李　俊　等著

清华大学出版社
北京

内 容 简 介

基于路由技术的 Internet，一个基本特征是网络行为的不确定性，其表现形式是通信路径和通信延迟的不确定性，随之而来的便是服务质量、网络安全、网络管理、地址分配等一系列先天性缺陷。本书提出了一种新的网络体系结构，将地址与拓扑结构相关联，以确定性的交换代替不确定性的路由，从而解决了骨干网络中的各种重大缺陷。书中系统地介绍了新体系结构中网络的拓扑结构、地址结构、数据包交换技术、通信流量均衡机制、短接通信技术、交换机或信道的快速自愈技术、服务质量控制方法、基于运营商地址空间与用户地址空间隔离的内建安全体系等。最后，本书介绍了新体系结构怎样促进网络向 IPv6 转换、怎样与现有的基于路由的 IPv4 或 IPv6 网络兼容、共存以及逐步向新体系结构过渡的过程。本书构成一个完整的体系，仅阅读部分章节不足以对新体系结构有完整的理解。

本书主要阅读对象是网络与通信领域的科研人员、高等院校教师和学生。

图书在版编目(CIP)数据

层次交换网络体系结构/钱华林等著. —北京：清华大学出版社，2008.
ISBN 978-7-302-17865-1

Ⅰ. 层…　Ⅱ. 钱…　Ⅲ. 计算机网络－网络结构　Ⅳ. TP393.02

中国版本图书馆 CIP 数据核字(2008)第 087906 号

责任编辑：丁　岭　徐跃进
责任校对：李建庄
责任印制：孟凡玉
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185×260　**印　张**：18　**字　数**：433 千字
版　　次：2008 年 11 月第 1 版　　**印　　次**：2008 年 11 月第1 次印刷
印　　数：1～5000
定　　价：39.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：(010)62770177 转 3103　　产品编号：028878-01

前　言

Internet是20世纪对人类社会最具影响力的科技发明之一。它为人类提供了全新的通信手段、信息交换手段以及信息获取手段，极大地加速了社会的信息化进程。目前，全世界已经有超过15亿人在使用Internet，并且每年还在以百分之二十左右的速度增长。发达国家的用网人口比例已经达到总人口的60%～80%，个别国家甚至超过了90%。在人口众多而网络建设起步稍晚的中国，前10年中上网人数平均每半年就翻一番，年增长率达400%，目前网络用户的数目超过了1.6亿。Internet这种迅猛的发展速度，远远超过了以往新技术如汽车、电话、电视等的普及速度。

Internet之所以发展迅速，是因为它顺应了社会向信息化进步的需要。回顾网络发展的历史，在Internet大规模发展之前，已经出现了很多采用不同通信协议并被大量使用的网络，有计算机公司研制的，有科研单位研制的，也有国际上联合作为标准而研制的。尤其是世界各国电信部门和厂商共同花费巨大人力物力制定的开放系统互连(OSI)标准，其精心设计的分层结构作为网络技术的典范出现在教科书中。Internet能够在众多的网络中脱颖而出，原因很多，但最重要的原因是协议开放、简洁，采用了包交换(packet switching)及路由(routing)技术。

OSI不能成功的一个重要原因是协议过于复杂，资料数量巨大、庞杂，难以读懂而且不能免费阅读，一个具有网络知识的专业人员，也很难在他(她)的有生之年读完相关的资料。协议的复杂性和协议文本的艰涩难懂使得技术人员难以方便地实现它，各厂商只能挑选其中自认为重要的部分加以实现，实现的功能不完备。这就造成同一厂商没有能力提供全套的网络协议软件，不同厂商的协议软件又难以做到良好的互操作性。为此，一些国家只好以政府的名义，在OSI协议中选取一个子集(称为GOSIP，government OSI profile)，让本国的厂商作为标准加以实现，以便国内各厂商之间的产品能够兼容。但这很难在采用不同GOSIP的国家之间互通。正是由于OSI协议复杂，实现成本高，协议软件十分昂贵，用户在经济上也难以承受。而TCP/IP协议却相反，它的协议十分简单，协议文本易读易理解，可以免费下载，协议实现容易，各厂商之间的协议容易做到互通和兼容。同时，这些特点使得很容易将协议软件作为操作系统的一部分提供给用户，用户不必单独为网络软件付钱。

包交换技术不同于以往的电路交换技术，它将用户要传送的完整数据分割成固定长度的小块，称之为包或分组，像一封信一样，加上必要的地址信息和顺序号码后在网络中独立传送，到达目的地后，再按照这些包的顺序加以组装，向收方投递完整的用户数据。在网络发展的早期，这是一项突破性的技术。一方面，当时的通信信道误码率非常高，大块数据一起

传，不出错的可能性很小。一旦出错，就要重传，而大块数据重传可能再次出错。小块信息的传输容易成功，个别有了错，重传的开销也小。更为重要的是，划分成独立包以后，不同用户之间的数据包可以穿插在一起传送，共享信道，大大提高了线路的利用率，节省了昂贵的信道费用。应该说，包交换是当时历史条件(信道贵、误码率高、速率低)下传输数据的最佳方案。

路由技术是与包交换技术相伴而生的。利用计算机的智能，在路由器之间定期交换信息，形成一张路由表。各路由器就能按照数据包封皮上写的目的地址，查看自己保存的路由表，将数据包一站一站地往目的地方向传送。所有这些转发工作，包括路由表的更新，都由路由器自动完成。早期发明路由技术的主要动机是希望每个节点有多条信道与外界相连，当出现自然灾害、战争等造成网络部件(尤其是信道)失效时，具有智能的路由器自动找到可通达的路，最大可能地保持网络的连通性。有了路由器对网络结构的自适应能力，人们还获得了任意接入网络的自由度，方便地组成任意连接的网状结构(mesh)。在网络部件特别是通信线路可靠性较低的情况下，任意连接的网络结构和路由器自动找路的方法是非常有效的。包交换和路由技术是 Internet 的核心和灵魂。

Internet 后来的蓬勃发展，则主要是网络应用的功劳。一方面，像电子邮件(E-mail)、电子公告板(BBS)、万维网(WWW)等应用层协议的出现，极大地吸引了用户上网；另一方面，传统的通信应用被移植到 Internet 上来，例如 IP 电话、视频会议、远程教育、视频点播等多媒体应用，使得 Internet 有取代其他通信手段的趋势。

然而，光通信技术的发展，使信道的带宽、可靠性和价格获得了极大的改善。当时针对信道低速、不可靠和昂贵所采取的技术措施，不仅没有必要了，反而成了影响网络通信效率的累赘。当网络规模不断发展，网络流量成指数增长，实时通信和通信质量保证的要求不断增长时，网络设备不堪重负；路由技术造成的通信路径不确定、同一对通信中双向数据包可能走不同的路径、大量的绕道、震荡和回路，使得对网络信道容量的需求不可预测，网络资源配置缺乏理论指导，信道利用率甚至比 PSTN 还低；实时通信要求服务质量控制技术(QoS)，而路由行为的不确定性，同一次通信中，数据包走过的路径随路由表的变化而动态地改变，使得无法沿通信路径预留资源，因而不可能真正实现 QoS；每天处理的 BGP 路由更新达百万次以上，峰值时，每秒要处理上万次更新，而每次更新，所有的核心路由器都要执行大量的通信和数据库更新处理；巨大的网络规模使得路由表达到 20 多万个表项，路由器不能快速有效地转发数据包；路由器越来越复杂，成本越来越高，耗电越来越多；慢速的端到端错误恢复手段，缺少网络自愈能力，使得 Internet 难以提供电信级的服务质量；任何一个核心路由器随时可能受到 DDOS 攻击，导致全网的不稳定；不良行为者可以仿冒别人的源地址而逃避对犯罪行为的追踪。所有这些困难，随着网络规模和应用需求的不断增长都变得越来越严重。再加上早期对 Internet 的发展规模估计不足，IPv4 地址空间将在两三年内耗尽。采用地址空间极大的 IPv6 协议，如果不考虑网络的层次化，路由表会进一步急剧增大，路由器的复杂性和它在网络中的瓶颈地位也就更加严重。目前大量的改进措施难以真正解决问题，陡然增加了网络协议的复杂性。Internet 早期协议简洁的优点也不复存在。Internet 到了必须进行重大改革的时候。

本书就是在分析了当前 Internet 难以应对的一系列困难的基础上，提出一种新的网络体系结构：层次交换网络体系结构。利用这种新的技术，彻底丢掉路由器；让网络回归简洁，设备简单、高效、便宜且降低功耗；让数据包的路径可预测，容易设计和配备信道资源并使信道

资源获得更加有效的利用；新体系结构下网络行为和数据路径的确定性，有利于真正实现QoS；网络结构信息和网络部件失效事件局部化，避免全局交换网络拓扑结构信息和失效事件，从而避免全局性的震荡、回路以及很长的路由收敛时间；让快速自愈能力不依赖于其他附加的网络技术（如SDH/SONET）；让多播树自然形成；让骨干网络设备的地址空间与用户地址空间隔离，像电话网络一样，用户无法干扰或攻击骨干网络设备；让源地址不可仿冒，有利于促进IPv6的部署和应用；等等。新的体系结构基本上克服了目前网络面临的各种困难。

任何想在Internet上做的改革或在Internet上使用一项新的技术，都必须与原有的Internet完全兼容、共存，否则将是不可行的。层次交换网络体系结构不要求现有网络作任何协议的修改就可以在一起工作。新体系结构可以逐步地、由小到大地部署，逐步替换旧的网络。这个特性使得层次交换网络体系结构是可部署的。

本书全面地介绍了层次交换网络体系结构及相关技术。由于对层次网络体系结构技术的研究是始终与原型系统、设备样机以及实验平台等工程项目并行进行的，除了原理性的描述外，书中列举了一些示例性的数据结构、表格，甚至实现方法。这些实现细节对读者更好地理解新体系结构，体会其简洁性和易实现性是有用的。对只要求粗略了解层次交换网络体系结构原理的读者，可以跳过这些示例性的细节。

层次交换网络体系结构的思想，是我们早在1999年开始对Internet路由行为的研究和观察并与成功的PSTN作比较后提出的，很多思路直接来自电话网路。但毕竟Internet的包交换与PSTN的电路交换有很大的差异。电话网络中，应用单一，只占用固定的窄带（64Kbps的PCM信道）；而Internet中的应用复杂多样，普通的电子邮件占用几Kbps的带宽就可以了，视频应用则需要几兆甚至几百兆的带宽。加之要考虑对现有网络的继承性和兼容性，不可能照搬PSTN技术。因而是一种全新的互联网络体系结构。这就注定了它会有较宽的涉及面，对诸如地址结构、交换方法、单点失效的避免、拓扑结构的灵活性、服务管理框架、QoS、网络管理、骨干网络的安全、源地址的不可仿冒性等一系列问题加以考虑，并且用原型系统和设备样机加以实现和验证。八年来，主要由硕士和博士研究生组成的研究小组成员，对体系结构的不断完善、对原型系统和样机的实现，做了大量的工作，在此对他们表示诚挚的感谢；他们是杨明川、马宏伟、牛广锋、鄂跃鹏、张道庆、周超、游军玲、吕红蕾、李晋、熊丹、代长城、李伟男、姜大伟、方蕾、林曼筠、娄雪明、申祥军……项目组先后归属于网络技术与应用研究室和中国科技网运行中心管理，在此对这两个部门的负责人南凯博士和张曦琼研究员的大力支持深表谢意。此外，项目组得到了来自各方面领导和专家的极大支持：中国科学院计算机网络信息中心前任领导阎保平主任、现任领导黄向阳主任和毛伟副主任，分别给予了所内经费的支持（项目编号：C30402，CNIC05001）；中国科学院路甬祥院长、胡启恒院士和戴博伟处长，为本项目争取了院长基金的支持（项目编号：KGCX2-YW-106）；国家发展与改革委员会支持、中国工程院主持的CNGI项目，为新体系结构提供了良好的实验、测试和试用平台；国家科学技术部863计划提供了项目经费（863项目编号：2007AA01Z214）；邬贺铨院士、邬江兴院士、刘韵洁院士、蒋林涛总工等一批资深网络专家，给予了极大的关切和支持，提出了很好的建议。在此一并表示感谢。

对书中可能出现的不完善和不正确之处，欢迎读者批评和指正。

作 者

2008年8月

目　录

第1章

Internet 面临的挑战

尽管 Internet 在过去的三十多年中取得了极大的成功，但这种成功完全是出乎意料的。当初设计 Internet 的体系结构时，有多方面的局限性，这些局限性给目前的 Internet 带来了致命的先天不足，使它面临着严峻的挑战，甚至威胁到它的可持续发展和生存。

早期的局限性主要体现在对 3 个方面认识不足，它们是：网络用途、网络规模和通信技术进步。创建 Internet 是为了满足军事和学术方面的需求，完全没有想到 Internet 会进入商用领域。对这些特定的用户群体，没有人认真考虑网络会受到群体内部成员的攻击或破坏，也没有人想到网络的赢利模式对网络的生存会有什么样的作用。由于对网络用途认识的局限，进而造成了对网络规模认识的局限。设计网络的先驱者们不但没有想到 Internet 会进入商业，也没有想到它会被普及到政府、机关、团体、家庭和任何个人，更没有想到它会延伸到个人、汽车和家庭等环境中的各种用具以及野外的数据采集与监测等领域。网络规模的迅猛发展和通信量的指数增长，对网络资源、网络性能和网络行为带来了极大的挑战，迫使网络设备不断复杂化、网络变得更加脆弱、服务质量无法保证、运营商对高昂的网络资源投资心存疑虑以及更为严重的网络安全隐患。另外，技术因素也严重影响了设计方案的合理选择。最初设计 Internet 时，通信技术相当落后，不仅信道容量小，更为严重的是信道的误码率高、信道的可用性差。数据通信对传输准确性的要求是百分之百的准确，而传统的话音通信是没有这种要求的。为了网络传输的可靠性和网络的可用性，不仅限制了报文的长度、设计了复杂的校验和重传技术，还创造性地提出了分布式的路由技术。利用计算机的智能和自学习能力，相互交换路由信息，能自适应地避开不可用的端口和信道，即使在遭到战争或自然灾害的局部损毁时，仍能将数据包送到目的地。这种路由技术的分布式特性，带来可用性好处的同时，也带来了极大的弊端。而当光通信技术日趋成熟后，误码率和可用性的问题并不突出，人们希望简化网络结构。例如以太网，为了可靠性和可用性，最初的设计是分布式的总线结构，任意一台挂在总线上的计算机失效，都不会影响其他计算机之间的正常通信。这种分布式结构带来的毛病是控制复杂、难以部署、难以管理和维护。在设备和信道可靠性获得大幅度提高后，人们很快就摒弃了分布式结构，采用树状结构的以太网交换机。但对 Internet 的路由技术和基于路由技术的无结构网络拓扑，人们意识到它的严重缺陷，进行了大量的研究，却拿不出有效的解决办法。

赫尔辛基大学的汉努·卡里教授多年前就指出，一个拥挤、设计糟糕和不安全的系统注定要消亡[1]。未来的互联网，要么被改变，要么就死亡。这并不是危言耸听，而是大多数对 Internet 体系结构、协议和网络行为有深刻认识的科学家的共识。美国的 GENI[2]、FIND[3]

和 100×100[4]等项目，都是针对当前互联网存在的严重问题而部署的研究课题。美国科学家认识到 Internet 日益严重的问题，甚至考虑采取与现有互联网完全不兼容的体系结构也在所不惜。中国工程院副院长邬贺铨指出，从修补式到革命式的路线仅仅是时间问题，互联网的发展已经到了十字路口，需要有一个新的起点。信息产业部的蒋林涛总工在一次会议上指出：已经看到互联网正在变糟，这是世界各国都认可的。目前信息基础设施非常容易受到预谋的攻击，可能会造成灾难性的后果。在美国形成了一个主流意见，就是要创新思考互联网基本体系结构，要采用新的设计理念[5]。笔者从 20 世纪 90 年代后期就看到了这些问题，并坚信在骨干网络中用交换技术代替路由技术是一个极其理想的解决办法。

当前的互联网究竟有哪些问题呢？主要在于：无序的网络体系结构、不确定的网络行为、不确定的网络延迟、局部故障全局化、不能保证通信服务的质量、安全问题无法解决、网络设备低效复杂昂贵、糟糕的网络可管理性、网络运行费用高、缺乏可持续发展的收费模式等。这些问题严重地阻碍了互联网的未来发展，严重地影响了网络运营商(ISP)对网络的可持续运行。

1.1 无序的网络体系结构

网络的体系结构主要包括：网络的拓扑结构、网络的地址结构、网络的协议功能结构 3 个方面。

网络的拓扑结构主要有树状结构(星状结构是它的特例)、环状结构和网状结构。前两种结构，节点间信道的连接规则是确定的：一根也不能多，一根也不能少。但在网状结构中，所有节点保持连通的前提下，信道可多可少，连接方式是任意的。对 N 个节点的网络，最多可以有 $N\times(N-1)$ 条信道。对一个小规模的网络，这种网状结构是不难控制和管理的，但当网络规模变得巨大，无人能说清楚网络是怎么连接的，新接入的网络想往哪儿连就往哪儿连，网络的跨度很大且很不均匀，节点的度数变成无尺度，这时候的网络就变成了一个无结构的、杂乱无章的网络，无法提出高效、合理的控制方法，只能用分布式方法对其进行控制和管理。而分布式控制意味着任何被控事件、控制信息和控制过程都是全局性的。对巨大网络系统，全局性的行为在网络资源的消耗、控制的可靠性、系统的可扩展性等方面都会遇到极大的困难。目前的互联网拓扑结构，已经无可挽回地陷入了杂乱无章的无结构状态。

地址的作用不仅在于指明要找的客体对象，更重要的是方便地找到编址对象。在我们的社会活动中，一个地名包含了国家、省市、县区、乡镇、街道和门牌号等信息，它不仅确定了一个家，还确定了怎样找到这个家的路径信息。图书馆里一本书的编号，不仅指定了一本书，还提供了怎样根据编号中的学科类别、子类、分类号等能在书架上很容易地找到这本书的信息。地址指定某客体对象的功能称作标识功能，地址表达怎样到达某客体对象的功能称为提供路径信息的功能。标识和路径信息本身都是静态的信息，获得了一个地址，只表明你拥有了某客体对象的标识和路径信息，并不意味着已经到达或得到了这个对象。到达或得到客体对象的过程是一个动态的寻路过程，在通信领域中称为路由过程(routing)。在网络通信技术中，常常把路径信息(path)与路由信息(route)混用(在中文里，把路由过程

routing 和路由信息 route 都翻译成路由，在不同的上下文中，路由一词，有时是名词，表示路径；有时是动词，表示寻路）。在互联网中，域名和 IP 地址都只有标识的作用，没有提供路由信息的作用。域名和 IP 地址都能标识一台计算机，前者使用字符，便于人的阅读和记忆，后者使用数字，便于计算机的处理。域名解析协议只解决两种标识之间的映射和转换，并不能提供任何路由信息。拿到一个 IP 地址，无论是人还是计算机，都无法直接从地址本身看出目的地在哪里。必须依靠一套互联网特有的路由协议才能获得路由信息。IP 地址的这种缺陷，造成了地址空间的无序性。这种无序性的另一个体现，就是从地址分配机构获得 IP 地址时，从技术上没有任何限制，同一个 ISP 的多块地址，可以分布在地址空间的任意位置。网络地址无序的根源在于网络拓扑结构的无序，地址结构本来应当与拓扑结构相关，但在现有的互联网中，由于拓扑结构的无序，两者是截然分离的。互联网中，为能把带有目的 IP 地址的数据包送到目的地，只能愚笨地建立一张庞大的路由表，为每个数据包查表后才能确定输出端口。为了建立路由表，要让所有网络转发设备互相交换信息，并记住去往所有目的地应走的输出端口。这种不能从地址中直接看出如何转发而必须通过查路由表才能知道如何转发的方法，曾被视为一大创新，现在成了一个致命的缺陷。

网络协议的功能划分，目的是希望实现网络功能时清晰、简单。通常的做法是，首先在通信子网和资源子网之间进行严格的分工：通信子网只负责把 IP 包从源节点送达目的地节点，所有其他工作都由资源子网实现。通信子网的设施由路由器、以太网交换机、集线器和通信链路等组成。资源子网则由用户的计算机（PC、服务器等）组成。但是，为了应付互联网的很多先天不足，在通信子网中修修补补，加入了很多本该在资源子网中执行的功能，造成了通信子网中协议功能的不清晰、复杂化和无序性。其次，在通信子网中，协议本应只包含物理层、数据链路层和网络层等的最低三层，但为了克服网络体系结构的先天性缺陷如复杂、低效和缺乏安全性等，加入了处理第四层甚至第七层协议数据单元的功能。在通信子网中加入高层协议功能，进一步造成通信设备的复杂化和低效率。MPLS[6] 的引入，目的是要克服路由带来的低效率，但却引入了奇怪的协议层次，它使得通信子网的协议层次，从三层（物理层、链路层和网络层）变成了五层（物理层、链路层、MPLS 数据交换层、网络层和 MPLS 路径控制层）。当 MPLS 转发数据包时，它在路由层之下；当 MPLS 生成交换路径时，它在路由层之上。但由于网络拓扑结构的无序性，不可能用新的基于包交换的网络层来替代原有的基于路由的网络层。这种在路由协议的上下各架设一层的奇怪做法，不仅不能克服路由协议造成的任何缺陷（因为它还在），反而增加了通信子网的复杂性。关于 MPLS 的局限性，将在下文详细分析。

对只有三层协议的通信子网而言，物理层和数据链路层都是局部的，即使是非广播多重访问（NBMA）链路（如 FR、ATM、X.25 等），从 IP 网络层看，它们都是一条可以连接两个或两个以上设备的点对点或多点链路，其行为也是局部的。但服务于端到端连接的网络层协议，因为两个通信端设备可能穿越整个互联网的最大跨度，因而是全局性的。在设计全局性的路由协议时，是针对无结构网络拓扑的，因而即使实际网络中确实存在某种结构，它也无法利用。这就在拓扑结构和路由协议之间形成了一个悖论：由于网络拓扑结构的任意性，网络层协议必须具有全局性；由于网络层协议的全局性，即使实际的网络有某种结构特性也无法得到利用。要摆脱这个死循环，只能两者同时作改变。如果再考虑地址结构的无序性，就会发现，拓扑结构无序、地址结构无序和网络层协议无序三者是互相关联的，三者是同