# 层次式交换网络—未来 Internet 的一种新框架

葛敬国　钱华林
（中国科学院计算机网络信息中心 NOC　北京 100080）

## Switched Hierarchical Network —a New Architecture for Future Internet

GE Jing-Guo　QIQN Hua-Lin
(Computer Network Information Center, The Chinese Academy of Sciences, Beijing 100080, China)
E-mail: gjg@cstnet.net.cn

**Abstract**　With the dramatic growth of the network scale and users, Internet is becoming a global information infrastructure. Driven by both new inventions in communications technologies and new applications, the architecture of Internet has evolved from the goals of openness and integration to those of high-performance, scalability and manageability. The paper first reviews the challenges faced in Internet and the existing solutions, and points out that the main reasons resulted in the bottleneck of core routers and the difficulties of QoS and performance guarantee rooted in the unstructured topologies and flat addresses structure of the Internet. The paper introduces the ideas of well-structured hierarchies into the topologies and address space of Internet, which takes advantages of the properties to construct the high-performance and scalable architecture for future Internet. In the end, we propose a model of switched hierarchical network which includes the extended topologies, critical protocols and control algorithms, and prove the correctness and feasibility of the model by network simulation experiments.

**Keywords**　Architecture, Hierarchical structure, Topology, High-Performance, Scalability

## 1　背景

新的应用需求和通讯技术的进步是驱动网络体系结构和网络协议发展的主要力量。自 1995 年开始，Internet 数据流量以每年超过 100%的速度增长，估计 2002 年，数据业务量将超过话音业务量[1]，并在未来三年里将占总业务量的 90%。同时，数据应用也从传统应用（电子邮件、文件传输、WWW 等）向实时多媒体应用（IP 电话、视频会议、VOD、交互游戏、虚拟现实、远程医疗和远程教育等）迅速发展，要求 Internet 提供大的通信容量和严格的 QoS 保证。最近十年，光纤传输技术，特别是 DWDM 被广泛应用于通讯领域，其传输能力可以用光 Moore 定律（每 6 个月光纤传输能力翻一番）描述，比描述集成电路发展规律的 Moore 定律（每 12—18 个月翻一番）快了 8 倍，能够满足应用需求以每年 2.3 到 2.5 倍的增长速度。尽管目前使用 DWDM 技术的一条光缆传输速率已经达到 Tbps 的数量级，但是现有的 Internet 网络远不能普遍地满足实时地支持视频、话音和数据通信的要求。问题出在两个方面：一是目前的路由和交换设备交换能力比光纤的传输能力差三个数量级，成了整个网络的瓶颈；二是目前的网络结构和协议体系，也不能适应保证 QoS 的大容量的数据传输要求。

未来网络的发展趋势是“更大、更快、更安全、更及时、更方便”，要求下一代网络体系结构具有开放、集成、高性能、可扩展、可管理等特点。Internet 经过几十年的成长，在开放与集成方面积累了很多成功经验和成熟技术，如自治的分布式结构、无连接的分组交换技术、网络互连协议 IP、端到端策略、网络地址独立于网络物理地址以及开发通用应用的技术[2]等，为我们研究新一代网络提供了非常好的参考。Internet 的设计背景是当时网络技术尚不成熟，网络体系结构丰富多样，互操作能力差，Internet 的设计目标是解决这些异构网络的互联问题。Internet 遵循这些设计原则保证了设计目标的实现，并获得了巨大的成功。但是随着应用需求，计算机与通讯技术、网络规模的不断飞速发展，Internet 进入宽带高速互联阶段，逐步成为承载视频、音频和数据等多种业务的统一基础通讯设施，要求 Internet 体系结构向高性能、可扩展和可管理目标迈进。原有的 Internet 体系结构设计条件已经发生改变，其中的一些原则在现有的网络技术条件下可能意义不大或者导致低效率。虽然为保证网络可靠性和互联的方便而使用无中心、无结构的拓扑设计具有健壮性和灵活性，但是随着网络规模膨胀和用户呈指数增长，将导致网络的性能、效率、QoS 保证、可扩展性、可管理性受到挑战，具体表现在复杂的路由计算、庞大的路由表和不确定的路由而引起的流量突发。Internet 界提出一系列方法和措施，旨在摆脱上述困境，但由于没有从根本上改变 Internet 拓扑与地址结构，结果是把系统越做越复杂，系统越复杂则带来处理时间越长、占用资源越多、效率越低，恶性循环，难以解决。现有的 Internet 体系结构设计限制了 Internet 的进一步发展，对现有 Internet 进行各种修补，事倍功半，要求从根本上改变 Internet 体系结构。

为了解决 Internet 面临的困境，我们在 Internet 拓扑结构、地址空间中引入层次结构的概念，使用层次式交换技术构造结构化的高性能、可扩展和可管理的网络。

## 2　Internet 面临的严峻挑战

当前 Internet 界研究的领域，主要集中在提高网络速度和容量（主要瓶颈在路由/交换设备），保障 QoS，扩展地址空

间(IPv6),增强网络可靠性、解决网络安全,开发各种高级网络应用等若干大的方面。其中前三个领域,都被 Internet 固有的设计缺陷所困扰,其它领域也因受前三个领域的牵连而增加了解决的难度。

### 2.1 路由器/交换机设备瓶颈问题

路由器的速度一直是人们关注的热点。路由器慢的根本原因在于查询庞大的路由表。随着 Internet 规模的不断扩大,全球的路由表项急剧膨胀。在 Internet 上执行 BGP 协议的路由器通常拥有数十万条路由表项,多路径选项和基于策略路由加剧路由表项的增长[3]。使用 CIDR 技术[4]推迟了 IP 地址短缺的严重程度和减小了路由表的长度,但需要最长匹配对路由表的查询带来了新的挑战。传统的路由器是用软件来负责选路和转发的,速度慢、延时大,无法满足网络发展的需要,为了提高路由器的速度,除了使用最先进的专用集成电路技术外,对路由器的体系结构作了精心的设计,从单总线、单 CPU 发展到多总线、多 CPU 和交换矩阵,采用 ASIC 用于包转发和路由表的查寻以及第二层交换与第三层选路的一体化;越来越多的功能放到接口卡上实现,以减轻中央 CPU 的负担;发明快速路由表查询算法;使用高速缓存技术;需要大规模的并行,将路由解析过程分步实施,达到流水作业的效果;需要复杂的互连,利用快速交换引擎或高速共享存储结构提高中央交换处理速度;将路由表的生成与路由表的使用相分离等等。但由于路由表庞大,无论怎样改进查表算法,为一个 IP 包寻找输出链路而进行的路由表查询,都是十分费时费事,运算量大,速度始终无法获得根本性的提高。特别是光传输技术的飞速发展,使得路由/交换瓶颈更加突出。路由/交换设备的相对低速,造成的影响有:减少了设备能处理的用户数量;增加了端到端的传输延迟;增加了队列的长度以及队列溢出引起的丢包数量;增加了端到端传输延迟的抖动度。所有这些缺陷,都不利于承载高速实时多媒体信息,更无法提供普遍的实时多媒体服务。为了提高路由器的速度,人们试图尽量用交换来代替路由,提出综合利用网络核心的交换技术和网络边缘的 IP 路由技术各自的优点而产生的 IP/MPLS 路由/交换标记结构[5]。目前基于此路由/交换结构的核心路由器的端口速度可达 10Gbps,但高速接口卡价格极其昂贵,而且,在 IP/MPLS 结构下,路由是永远避不开的,只能做到部分替代,并没有从根本上解决问题。第一,复杂的路由表生成过程、保存庞大的路由表及对路由表费时的查询工作没有消除;第二,如何有效地识别各个独立的、时间上随机到达的 IP 包的流类别。同时标记交换技术也引入了复杂性,例如标记的生成、分配、管理、标记到端口的查询和映射,标记交换与其它协议的关系等。Internet 界支持保持 Internet 骨干网的简单性和非智能性,将复杂性和智能性放在网络边缘和客户端,但 MPLS 则与这种方法大相径庭。MPLS 增加了 Internet 核心的复杂性,大大偏离了 Internet 结构。

完全依靠电子技术的路由/交换设备,难以较大幅度地提高交换速度。人们越来越多地把希望寄托在光交换设备上。目前成熟的基于波长粒度的光交换开关,类似于电路交换,线速度可以做得很高,但其交换机制仍然是机电式的,并不能满足 IP 交换的要求。纯光的基于包突发技术的交换设备,离实际使用还有很长的路要走。

### 2.2 QoS 模型和机制存在的问题

Internet 最初的设计目的是进行高效的数据传输,因此所使用的 TCP/IP 协议族是一种无连接的、基于数据报的传输模式。IP(IPv4)所提供的是一种"尽力而为(best-effort)"的服务,无法保证吞吐量和传送时延等服务质量。提供服务质量保证是 IP 骨干网提供实时、多媒体通信和关键应用等多业务的必要条件。IETF 已经建议了很多服务模型和机制,以满足 QoS 的需求。其中比较有名的有:IntServ/RSVP 模型[6~8]、DiffServ 模型[9]、基于多协议标记交换(MPLS)的流量工程[10]和约束路由(Constrainted Based Routing)[11]。IntServ/RSVP 模型的特点是资源预留,实时应用在传输数据前必须用 RSVP 预先建立通道和预留资源。IntServ 尽管能提供 QoS 保证,但扩展性较差。因为其工作方式是基于每个流的,在骨干网上,业务流的数目很大,同时它还要求路由器的转发速率很高,不能对每个包执行复杂的调度,这使得 IntServ 难于在骨干网上得到实施,只可能应用于边缘网络中。由于网络中传输的包流量分布的随机性和突发性,这种企图对无序的交通加以有序管理,在无连接的网络中勉为其难地加入面向连接的全程信令的想法简直是一个悖论:为了在资源上保证传输的质量,需要复杂的算法和协议,而这些算法和协议反过来与用户流量争夺宝贵的 CPU 和信道资源。近年来,QoS 模型研究转向简单、可扩展的区分服务(DiffServ)体系结构。虽然其实现简单,但缺点是难以提供严格的端到端 QoS 保证。DiffServ 模型本身还不完善,DiffServ 的实现机制和性能分析仍然处于研究之中,包括业务类别的具体划分、每类业务性能的量化描述,因此利用 DiffServ 模型实现对 IP 网络 QoS 的保证目前尚不成熟。流量工程是一种安排通信流量如何通过网络的过程,以避免不均匀地使用网络而导致拥塞的过程,实现网络资源性能优化,是对 DiffServ 模型的补充。为使流量工程自动化,约束路由和 MPLS 是重要的工具。MPLS 与约束路由相结合,利用约束路由为流寻找一条最大满足 QoS 需求和其它策略限制的最优路由,MPLS 显式路由功能根据约束路由或手工指定的路由通过带有 QoS 参数的信令协议建立受限标记交换路径(CR-LSP)[12],因而能够有效地实施大范围的流量工程,实现负载平衡避免拥塞、故障或者提供特殊类型的服务。但约束路由与 MPLS 本身还不成熟。

造成上述 Internet QoS 模型实施困难的根本原因在于 Internet 自身结构庞大而复杂,至今没有一个统一的理论模型可以分析和预测网络的动态变化过程,难以预测资源分布和流量分布,使得在动态的环境中为保证一个流的 QoS 而寻找、建立并维持满足资源要求的路由非常困难。对于简单网络,可以让网络管理员手工配置链路的代价,均匀地分配流量是可以的,但对于复杂网络,这几乎不可能。现在的动态路由协议 RIP,OSPF 和 IS-IS 使用最短路径路由算法,总是选择最短路径转发包,即基于单一测度(metric)、管理权重或跳数决定最优的包路径,首要目的是保证基本的连通性,都会导致不均匀的通信分布。OSPF 的等价多路径(ECMP)选项,在给多个最短路径分配负载时是有用的。但是,如果只有一条最短路径,ECMP 就无能为力了。Internet QoS 实现模型或机制如 IntServ/RSVP、DiffServ、流量工程和约束路由都需要实现对有 QoS 需求的流进行带 QoS 需求约束的路由选择机制,能够动态确定可行路径、优化资源利用率和网络性能。由于 QoS 路由存在多个 QoS 约束目标,QoS 路由问题的核心就是实时地对网络多约束条件下路由选择中的 NP 完全性问题求解。由于 Internet 无结构的特性导致 QoS 路由计算的复杂性,使得 Internet 的 QoS 模型仍然处于研究之中,难以从根本上得以解决。

### 2.3 IP地址空间扩展问题

随着Internet用户的迅速增加，网络地址不足的危机日益严重。按目前入网主机的增长速度预计，到2005年左右，IP地址将被耗尽。虽然采用内部地址的NAT(Network Address Translator)技术[13,14]、无类别域间路由(classless inter domain routing，CIDR)[15]技术暂时缓解了IPv4地址空间的不足，但移动IP、信息家电的应用将导致新一轮IP地址的短缺，因此有必要在互联网上引入新一代Internet协议——IPv6[16]。从地址空间的扩展看，用16字节的IPv6代替只有4字节的IPv4，地址是够用了，但从IPv4向IPv6过渡的实施难度，被资深网络专家喻为"给飞行中的飞机更换发动机"。目前可用的过渡方法有两种[17~19]，一种是双协议栈；另一种是所谓的隧道技术。隧道的方法虽然简单，但IPv6的计算机难以与IPv4的计算机通信。如要求这些计算机也能与数量巨大的IPv4计算机通信，它们必须同时也运行IPv4，或要求IPv4的计算机也能识别IPv6的报文，前者省不了IPv4地址，后者要求整个Internet的计算机修改协议软件；而如果一个站点具有IPv4/IPv6的双协议栈，它就既可以与IPv4站点通信也可以与IPv6站点通信。同时，上层协议也可能需要替换。但是这样做显然需要巨大的投资，特别是对于规模较大的网络。

造成上述问题的根本原因是最初设计Internet时对它的规模、应用范围没有足够认识而造成的。如果对Internet无结构的拓扑、平面地址结构和基于路由表动态选路等技术不加以彻底改变，只在这种有根本性缺陷的系统上修补，其结果是，把系统越做越复杂、付出巨大代价最终仍然找不到满意的解决方法，难以满足未来Internet的需求。

## 3 层次式交换网络思想的提出

传统的电话网络采用的是层次结构，其地址(电话号码)也相应地采用层次式结构，由国家代码、区号和本地号码组成，不仅作为电话机的标识，而且由于其地址结构与网络拓扑结构相匹配，可以表示其地理位置，使得路由非常简单，简化了交换机的复杂度，降低了成本。实践证明，电话系统具有非常好的可扩展性，随着用户数量和网络规模(链路速度和交换机容量)的不断进步而逐步升级系统，但系统的基本结构一百多年保持不变。实际上，现有大多数的包交换网络拓扑也采用层次结构，分为核心层、中间层、和接入层等层次。但是由于地址空间是无层次、一维平铺的，网络地址仅用作识别不同网络的标志，不可能与网络的拓扑结构相匹配，不能利用其层次拓扑的特点，只能通过查找庞大的路由表确定路由，造成转发瓶颈。Internet的设计者意识到这一点，采取了许多补救措施，在IPv6中地址分配使用基于汇集的层次地址分配方案，骨干网实际上大都使用层次拓扑结构，在一定程度上缓解了可扩展性压力，但是由于拓扑结构与地址结构相分离以及需要保证和原有历史遗留系统的兼容性原因，效果并不明显。

层次式交换网络设计思想主要体现在网络的拓扑结构是按层次结构构造的，网络的地址空间也按层次结构分配，并且拓扑结构的层次与地址结构的层次是严格匹配的。将层次结构引入网络的拓扑结构设计、地址空间分配，充分利用层次结构本身固有的可扩展性、可管理性和效率高的属性，简化路由和寻址功能，避免其不可靠性和不灵活性，解决未来网络的高性能、可扩展性、可管理性问题。

### 3.1 层次式交换网络体系结构是可扩展网络的必然选择

早期设计Internet时，一个核心的思想就是要保证可靠性。这样做有充足的理由：当时的信道失效率和误码率都很高；希望在自然灾害或战争环境下网络有很好的可存活性。他们从可靠性出发，各种设计都采取了无中心、分散的思想。IP的非连接特性、动态路由算法和高度分布式的网络体系结构都有助于实现网络的高度抗干扰性。Internet网络的拓扑结构，是任意连接结构，没有中心，没有层次，只要每个路由交换设备有两条或两条以上信道，就认为可靠性获得了很大的提高。IP地址的安排，以网号为基础，辅以网内主机号。而在路由算法中起决定性作用的网号是一维平铺(Flat)的，没有任何结构，没有层次性。对一维平铺的IP网络地址，当其数量十分庞大时，既不能设计出高效的路由信息交换算法，也无法设计出快速的路由表查询算法。这就是路由问题的症结所在[20]。其实，任意连接的结构和一维平铺的地址空间并不适合人类的思维模式和计算机高效处理的要求，更缺少可扩展性，只有在非常小的系统中才能以其简易而获得应用。

随着Internet逐步成为未来全球基础信息传输平台，用户数量、网络规模呈指数增长，新的应用需要提供有QoS保证的大数据量实时多媒体传输服务，原有Internet无结构的缺陷对网络规模扩展、网络性能扩展的影响越来越突出，成为制约网络发展的关键因素，必须在现有的Internet中引入结构化概念。人类社会的组织结构和地址空间，大多是层次式的，如政府组织、电话系统、邮政地址系统、Internet的域名系统等。由于层次式结构具有良好的可扩展性和可管理性，适合有效构造和管理大规模的网络，成为未来Internet首选结构。当网络的规模扩大，用户数量或用户网络增加时，其扩展只需增加中间层次即可，影响局限在局部范围内，不会对整个网络产生影响。

### 3.2 高性能的层次式交换网络的交换机制

层次式系统的一个显著特点就是它具有树形结构。计算机对树形结构的查询和处理速度为O(logN)，远比处理一维平铺的O(N)快。层次式交换网络的拓扑结构是按层次结构组织的，其地址空间也是按层次方式分配的，由于地址空间和拓扑结构严格匹配，如果将IP地址划分成不同层次的域号，各IP地址子域与网络层次结构相关联，IP地址本身包含了路由信息，处于树的两个节点间具有唯一固定的最短路径。IP报文从源节点沿这条唯一路径送到目的地，路径上的节点设备只要根据报文地址中特定的子域号选择输出端口，就完成了路径的选择。层次交换技术按照IP子域字段进行的局部控制的简单交换，不依赖于查询庞大的路由表以及复杂的全局性的路由技术，不再具有传统意义上路由选择的含义，路由器退化成了交换机。交换的依据是IP包的目的地址和源地址的相应字节(即相应的子域)，与标记交换相比，不需要分配、管理、检索很长的标记。层次式交换技术简化了交换机的复杂度，消除了复杂的路由计算和查找庞大路由表任务，可以用硬件完成大多数交换功能，提高交换机的性能，大大减轻路由器瓶颈。层次交换控制具有局部性，交换控制信息局限于交换节点内部，无须在整个网络传播控制信息，无须考虑与其它节点的互操作问题，不仅简化了交换机的设计，而且简化了网络控制协议，大大降低了网络协议的复杂度。复杂度的降低是构造高性能网络的关键。

### 3.3 局域自治的可管理网络

从管理的角度看，层次式结构本身是一种分层的自治系统，各层节点管理范围明确，限定在一个有限范围内，处在某

一层上的一个节点，其直接管辖范围为本节点内部事务及直接属于它的下层节点地址分配。下层在自己的管辖范围内所做的变动，上层可以不知道，局部网络的变化对网络其它部分影响不大。这样使得网络的管理负担分散在全网的各个层次的节点中，没有全局的管理，实现网络管理的局域自治，简化网络管理复杂性。将IP地址子域与网络层次结构相关联的另一个好处是，IP地址的分配失去了全局意义，各层自由决定下层的结构和地址分配，不再需要全球性的地址分配机构，不再出现占用地址的不公平现象。

### 3.4 层次式网络是解决超大规模网络QoS的根本途径

目前任意连接的网络，一条信道被不可知的用户群体借道使用，加上通信的突发性，设计人员无法正确为每条信道及其相应的端口分配合适的资源数量。网络的流量工程难以进行，或者方法复杂占用过多的资源，或者效果不好，造成网络负载的不均衡现象，难以保证QoS。层次式的网络结构，可以把子区域内局部通信的流量限制在子区域内部，不占用子区域以外的网络资源，避免了目前经常出现的混乱、低效、浪费网络资源的绕道现象。层次式网络有严格的拓扑结构和相对固定的路由，可以简化描述网络动态变化的网络模型，降低网络资源管理和网络流量分布预测的复杂性，提高流量工程性能，为QoS机制的应用提供根本保证。

## 4 层次式交换网络模型

层次式交换网络模型的核心内容，主要研究层次交换网络体系结构、关键协议和控制算法。

### 4.1 层次式交换网络体系结构的设计

层次交换网络引入层次结构，使网络具有鲜明的特点，主要研究层次式交换网络的拓扑结构设计，可靠性和灵活性是设计成败的关键。层次交换网络继续使用现有的Internet功能框架，只是考虑了层次结构对路由与交换的影响而简化了网络层路由选择和转发操作，其它各层保持不变。

根据管理域的不同，可以把层次交换网络分为两大部分：骨干网和接入网。骨干网由运营商运行和管理；接入网由用户单位管理，如图1所示，相当于把所有的用户接入网用一个大交换机连接起来。这个大交换机代替了原来由路由器组成的任意连接、复杂的、难以管理的骨干网。可以分别针对骨干网和接入网的特点设计各自的网络拓扑结构、地址结构、交换方式等。

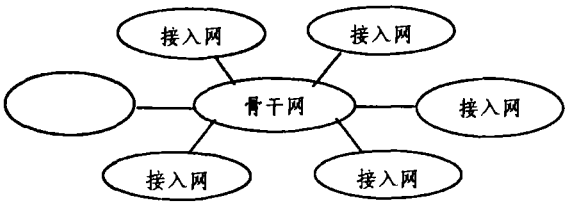


图1 网络的两层逻辑结构

在层次网络中，骨干网结构是树形结构，由多个交换层次组成。由于传统的简单树型结构从上层节点到下层节点的链路是唯一的，存在可靠性、容量可扩展性和灵活性等固有的缺点。一条信道或节点的失效，使得该信道或节点下连的子树失去连接性，造成一批下属子树的断连。单根信道的速率和单个交换机的容量受技术的限制，需要扩充扩展信道数量和核心节点容量。由于树型结构任意节点间的路径是唯一的，不能满足多个管理实体间灵活的路由策略需求。在设计层次结构网络拓扑时，需要扩展严格的树型结构，避免这些缺点。

1)基于逻辑节点域和逻辑信道的扩展

一种解决可靠性和容量可扩展性的方案基本思想是，以逻辑节点(或称节点域)替代树型结构的节点，以逻辑信道替代树型结构的分支，以逻辑节点和逻辑信道为基本元素按树形结构组织成逻辑上的层次结构网络，保持层次结构特点。从物理实现角度看，每个节点域中有多个交换机互联保证了交换节点容量的可扩展性和路径选择的灵活性，每条逻辑信道包含多条物理信道扩展了逻辑链路的带宽。逻辑扩展方法既保持了树型结构的特点，又保证了可靠性和可扩展性。

一个节点域由一批互相连接的交换机和其它功能服务器(如策略服务器、带宽代理服务器(Bandwidth Broker)和各类管理信息库MIB等)组成的地理位置分散的自治域，如图2。从逻辑上看，节点域是树形结构上的一个交换节点，其功能如同一个交换机，每个节点域有一全局唯一地址，下层节点地址的前缀包含上层节点的地址，完成流量的向上和向下两个方向的转发。

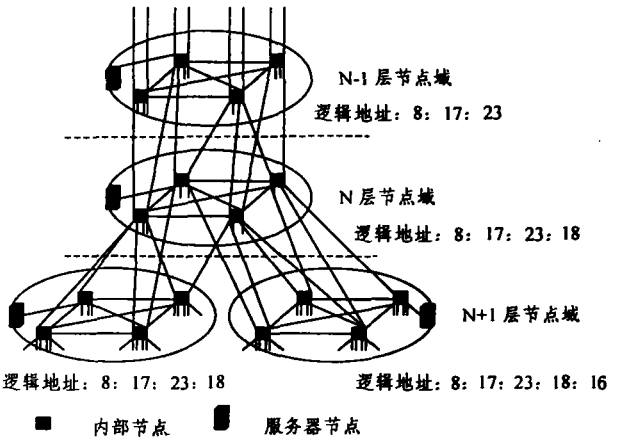


图2

与一个节点域相联系的有三类信道，即上行信道、下行信道、内部信道。将连接节点域的信道抽象为逻辑信道，每个逻辑信道由多条物理信道组成。对于某一节点域而言，有一上行逻辑信道和一组下行逻辑信道。上行逻辑信道只有一个，连接其父节点域，由若干物理信道组成，对应于每根信道的端口，既可属于域内同一交换机，也可属于域内不同的交换机。下行逻辑信道可有多个，连接多个子节点域，每个下行逻辑信道分配一个下行逻辑信道号。逻辑信道号其实就是下一层子域的子域号，由上一层节点域管理和分配。内部信道连接域内各交换机，在节点域边界外不可见。

由于层次式交换网络设计目标为构造大规模核心骨干网络，节点域设计目标主要考虑节点域的交换容量、转发效率、流量的负载平衡和节点可靠性。节点域设计主要研究节点域内部的物理节点和链路连接与配置、交换机制和交换性能，如信道选择机制、负载平衡能力、故障检测/恢复能力、域吞吐量和延迟性能以及各物理节点内流调度和缓冲管理策略等。节点域是一个由多个内部节点组成的自治域，内部可扩展，内部拓扑结构和交换机制对外是透明的，各个节点域内部可以使用不同拓扑结构和交换机制，由该结点的管理者选择。对于核心交换层，节点域内部结构、交换机制不宜过分复杂，否则影响包通过该层的延迟时间，应该在可靠性、容量、效率和代价间进行折中。

2)基于短接的扩展

另外一种方案主要解决严格层次结构的灵活性和效率问题,在不同子树间增加直接链路从而使得子树间存在第二条有效路径,子树间流量不一定需要它们共同的祖先节点转发。此种方法虽然保持了拓扑的层次性,但破坏了树型结构,需要遵循一定的扩展规则,保证路径选择和交换的简单性与层次性特点,在灵活性、效率和复杂性之间折衷,否则就容易变成任意连接的结构。为引入竞争,层次结构网络中存在多个ISP,需要考虑不同ISP网络间互联的灵活性和可管理问题。存在两方面问题:一是多宿主问题,用户节点为了容错、负载平衡或竞争考虑可以同时与属于不同ISP的多个上层节点域相连,如图3所示的C节点域同时是属于ISP1的A节点域和属于ISP2的B节点域的子节点,这样节点域C及其子树有两套逻辑地址,地址前缀分别为A、B节点域地址,使用不同的地址出入节点域C及其子树有不同的逻辑路径;二是不同ISP间流量交换的效率问题,严格的层次交换结构要求跨ISP的流量必须通过上层节点转发,即使同一地理区域的流量也不例外,造成资源浪费。如果两个节点域间流量非常大,可以直接在它们之间增加物理链路(称为直达主干),它们之间流量无须经过上层节点而直接到达,如图4所示,地理位置相同、分别属于ISP1和ISP2的A、B节点域有直接链路短接,逻辑上相互成为对方的子树,两子树间流量在局部进行交换,无须经过顶层接点转发,提高了网络效率。基于短接扩展的缺点是增加网络拓扑和交换机路由的复杂性,需要遵循严格的扩展规则和网络层交换协议,保证层次交换的特点。

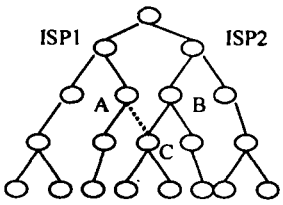


图3 多宿主扩展

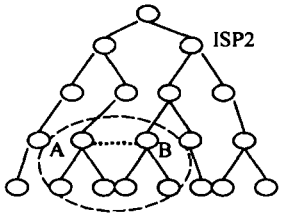


图4 对等短接扩展

3)顶层节点域设计

顶层节点域是不同国家、不同ISP间流量的公共交换节点,其设计需要满足三个方面的需求:管理上的公平性、连接的自由和灵活性、节点容量的可扩展性。管理公平性体现在节点域的管理权限和流量交换的对等、自由和公平,客观上需要满足属于不同管理域的网络间能够自由连接的拓扑结构,连接主要考虑策略因素而不是依赖技术原因。顶层节点交换容量巨大,为扩展性需要,需要考虑应用最新网络技术,特别是充分结合当前的光网络技术的特点。

4.2 网络层协议设计

4.2.1 地址分配及包结构设计 网络地址按照层次结构分配是层次交换网络的特点之一,将地址划分为多个子域,每个子域对应不同的层次,各层独立管理和决定自己的地址空间分配。层次网络中的IP地址,采用与IPv6兼容的方式,使用16字节。把层次网络的16个地址字节分为两部分:前8个字节用于骨干网,后8个字节用于接入网,如图5所示。骨干网采用基于汇聚的层次式地址分配方式为层次式交换结构奠定基础。接入网的网络地址前缀由其接入点骨干网地址确定,网络地址的后8个字节分配由用户单位负责,接入网内部可以使用与骨干网同构的地址分配和交换结构,也可以使用当前的地址结构和路由方式,在向全层次式网络过渡的过程中,接入网可以有不同的网络结构同时存在。

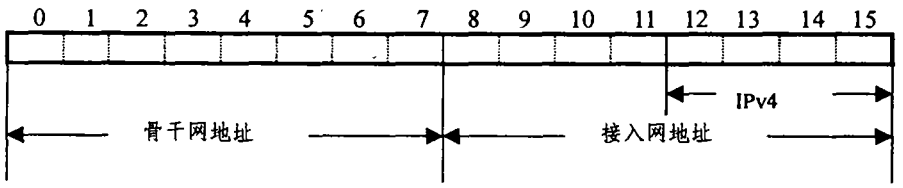


图5 骨干网与接入网IP地址的分配与管理界限

骨干网层次地址的划分有两种方案:基于地理区域划分和基于管理区域(如国家、ISP)划分。地理区域划分一般网络效率高,但对于一个跨区域的ISP存在管理问题。管理区域划分可能导致路由迂回现象。一般结合两种分配方案,在不同层次根据实际需要侧重不同方案。

层次交换网络包头结构采用IPv6包头格式。

4.2.2 交换协议设计 交换协议以节点域为基本单位,独立于其它节点域,在每个节点域内部实施。因此,不需要在全网进行整个网络拓扑结构及路由信息的传递与交换,保证协议的简单性。交换控制协议设计要解决的问题只有两个:1)怎样控制和管理一个逻辑节点域内的各个交换机,使其对外像一个交换机一样,对内有效、合理地发挥每个交换机的作用;2)怎样控制和管理一条逻辑信道内的各物理信道,使其对外像一条信道一样,对内均衡地在各不同速率的物理信道上分担通信量。交换协议主要完成两部分功能:逻辑节点域内部控制信息交换与路由计算和数据包路由选择。数据包路由选择分为两部分:逻辑信道选择和物理链路选择。

为了管理各交换机及其端口和信道的工作状态,各交换机之间必须有内部的通信,节点域内部交换的控制信息主要有交换机配置表和交换机信道状态表。各交换机互相交换了交换机配置表后,各交换机计算动态地生成域配置表,用以描述域内拓扑及信道状态,交换机的连接及其端口的类型、速率、信道号的分配等,各交换机保存的域配置表内容是相同的。在各交换机之间交换了信道状态表后,就可以生成域状态表。域信道状态表描述各物理信道的实时通信情况,内容为端口的速率,当前的队列长度,空余缓冲区大小,各交换机都保存一致的域信道状态表。信道状态表的传送,要及时反映各对外物理信道的忙碌情况,由定时器启动,时间间隔应该很短,例如毫秒数量级,以尽快反映信道的当前状态,供调度、均衡各物理信道负荷使用。

包的源地址和目的地址对唯一确定了其逻辑路径,包到达一个节点域时首先进行逻辑信道的选择。对上行到达的报文,一种可能是继续向上一层传送,另一种可能是改为下行。改为下行的判断条件是:目的地址和源地址在本层及其以上各层对应的地址字段完全相同。下行报文按目的地址中对应下层地址字段选择下行逻辑信道号。当存在短接扩展时,逻辑

信道的选择需要重新设计。

选择了逻辑信道后，需要进行物理信道的选择。物理信道的分配，解决两方面的问题：一是在各物理信道之间均衡分配通信量；二是处理物理信道的失效。有两种方式分配物理信道：一种叫做按流分配，要求同一流（对应于同一源/目的地对、具有某种相同特征的通信数据包序列）的数据包序列走相同的物理信道，保持每流中各包的次序不变；另一种叫做按包分配，不考虑每个 IP 包与流的关系，以包为单位根据各物理信道的状态对包作独立的物理信道分配。按流分配存在两个问题：一是流的识别分类问题，没有快速有效算法；二是流在物理信道上均匀分配问题，由于不同流的流量特征变化很大，基于流为单位的负载平衡粒度较粗。按 IP 包分配，需要解决物理信道状态信息的实时交换和为每个包进行物理信道选择的效率问题。

## 5 交换控制算法研究

层次交换网络协议的核心内容是各种控制算法，如准入控制算法、拥塞控制算法以及信道选择控制算法。这些协议算法是决定协议是否可行以及性能是否优劣的主要因素。这里研究信道选择控制算法。

### 5.1 基于按流分配的交换控制算法

要实现或近似地实现按流分配，可以有如下四种办法：随机数分布法；Hash 分布法；交换机记录逐流状态分配；数据包携带信息分配。

1）*随机数分布法* 基于随机数分布的方法，希望对流的参数或状态不作任何记录，只根据到达 IP 包特有的某些参数（通常用它的源地址和目的地址对）作为种子进行随机函数计算，获得一个与某条物理信道相对应的数值，进而选定该物理信道。由于从属于同一个流的各数据包提取的种子相同，随机函数的计算结果总是相同的，故一个流中的各个包都会选择相同的物理信道。

2）*Hash 分布法* 此方法与随机数分布法类似，但考虑了流的带宽需求，更具准确性和灵活性，适合进行流量工程。一般地，当信道的带宽利用率达到一阈值时，就需要考虑做流量工程，重新为 Hash 表各聚集流分配信道。Hash 表大小选择很重要，Hash 表太大，存储、管理开销大，太小则聚集流粒度较大，信道选择缺少灵活性，流量工程效果不佳。

3）*交换机记录每流状态分配法* 最简单的办法当然是让沿路的每个节点记录每个流的状态和参数，一个包到达时，查表确定该包属于哪个流，应走哪条物理信道。但和 IntServ 一样，让每个节点记录并查询大量的流状态，系统开销太大，缺乏可扩展性。这种方法只适用于很小的网络系统或大型网络的边缘节点。

4）*数据包携带信息分配法* 此方法与交换机记录逐流状态分配的方法类似，对每个流的分配情况加以记录。不同的是，不让中间节点来作记录，而是让边缘节点来记录。每个流建立时，沿路除分配资源外，还填写分配的物理信道号，边缘节点将其记录下来。对进入网络的包进行封装时，填写沿路物理信道的信息。当包进入节点域，根据包中携带的信息选择物理信道。不足之处是增加了包的长度，增加了网络资源开销。

5）*四种方法的比较* 随机数分布方法比较简单，因为不必作任何状态信息的记录。这种分配的缺点是，没有考虑各个流的带宽需求，使得各物理信道分配的负荷有可能不够均衡，分配的合理性在很大程度上取决于随机函数的性能。由于源地址和目的地址中的不少字节是相同的、有规律性的，很难真正得到随机均匀分布，使得信道分配不够均衡。Hash 分布法比随机数分布法灵活、准确，Hash 计算将数量巨大的微流汇集为有限数目（Hash 空间 K）汇集流，保证汇集的均匀性，考虑每个微流不同的带宽需求预留汇集流带宽，基于汇集流为单位的流量分布依赖于实时的流量工程。交换机记录每流状态分配法和数据包携带信息分配法都可以做到合理、均衡的分配，前者缺乏可扩展性，后者将增加数据包的开销，降低通信效率。实际上 Hash 分布法可以与交换机记录每流状态分配法相结合，采用 Hash 汇集方法解决记录流状态的扩展性问题。

### 5.2 按数据包分配

按数据包分配有两种方法：加权循环法和基于可用度分配法。

1）*加权循环法* 在逻辑信道的所有物理信道中用加权循环的方法为到达的 IP 包选择物理信道，每个信道的权值为其带宽占逻辑信道总容量的比例。

2）*基于信道可用度分配法* 信道可用度算法实时收集链路状态，用一评价函数计算逻辑信道中各物理信道的资源可用度，以此标准为进入域的包选择物理信道。

## 6 初步实验结果

我们扩展 NS-2 模拟构造了一个基于节点域扩展的层次式交换网络。实现了交换协议，验证了依赖层次式交换网络的层次拓扑结构和层次地址结构以及二者匹配关系进行数据通信的正确性。初步比较各种交换控制算法的负载平衡能力和计算复杂度。实验结果见文[21]。

**结论与未来的工作** 本文分析了当前 Internet 面临的挑战和其主流解决方案的关键技术，指出 Internet 任意结构的拓扑、一维平铺的地址结构以及拓扑结构与地址分离的特性对于构造大规模的全球基础通信平台存在致命的缺陷。本文在现有 Internet 体系结构中引入层次结构思想，充分利用层次结构本身固有的属性，简化路由和寻址功能，解决未来 Internet 的高性能、可扩展性、可管理性问题。本文最后给出了未来 Internet 骨干网络框架—层次式交换网络模型，提出各种拓扑扩展模式、网络交换协议和交换控制算法。

下一步的工作主要集中于进一步完善层次式交换网络的拓扑扩展、节点域内部结构设计和交换协议设计，实现、部署层次式交换实验网络，对其进行网络性能分析，以及考虑与现有网络的兼容问题。

## 参 考 文 献


1 Coffman K, Odlyzko A. The Size and Growth Rate of the Internet. First Monday. 1998, 3(10)

2 Carpenter B. Architectural Principles of the Internet. Request for Comments (Informational) 1958, Internet Engineering Task Force, June 1996

3 Huston G. Commentary on Inter-Domain Routing in the Internet. Internet Draft, draft-iab-bgparch-02. txt, Sept. 2001, work in progress

4 Fuller V, Li T, Yu J, Varadhan K. Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy. RFC 1518, Sep. 1993

5 Rosen E, Viswanathan A, Callon R. A Proposed Architecture for MPLS. RFC 3031, Jan. 2001

6 Braden R, Clark D, Shenker S. Integrated Services in the Internet Architecture: an Overview. RFC 1633, ISI, MIT, and PARC, June 1994


（等静第10页）

网络元素进行自适应行为协商、达成一致的交互方案。主要是自适应 QoS 机制协商协议的设计。例如：接收端节点由于 QoS 降级要求数据的编码方案发生变化时，可通过信令机制与发送端协商，使发送端和接收端达成一致的数据编码方案。

· 行为机制　自适应 QoS 机制所实施的行为，例如：资源协商、端对端的语义协商、对应用程序参数的修改等。自适应 QoS 行为机制可能包括一系列的行为。行为机制由触发机制触发。对于不同的资源变化情况，有不同的行为实施。

· 触发机制　实现自适应 QoS 行为机制的触发。

性能问题是要考虑的主要问题。检测机制和触发机制可能发生在一个端节点上，而自适应 QoS 行为机制是一系列的行为，如资源协商、端对端的语义协商、对应用程序参数的修改等，涉及到两个端节点和可能的网络中间节点。不适当的自适应行为：(1)对网络产生一定的负载，并占用一定的网络资源；(2)可能造成应用程序的不稳定。所以触发策略的设计对自适应 QoS 机制是非常重要的。如何过滤"细粒度"变化，是触发机制要重点研究的问题。

· 检测机制　实现网络资源的检测，为触发机制提供决策依据。检测机制主要应考虑：(1)检测点；(2)检测时间周期；(3)检测参数。

基于发送者驱动策略的自适应 QoS 机制的检测点应包括发送端到接收端所经过路径的所有网络中间节点。基于接收者驱动策略的自适应 QoS 机制的检测点只在接收端节点本身。对于无线网络，由于资源的变化主要体现在无线网段，一般应选择基于接收者驱动策略，只在接收端检测。

检测时间周期的确定影响系统的效率。检测周期越短，占用系统资源越大，但检测周期过长，影响资源变化的准确判断。

检测参数选择与具体的应用相关。应选择影响应用的主要参数进行检测。

实现自适应 QoS，要求应用是：

(1)具有 QoS 感知，同时具有 QoS 软保证(soft QoS guarantee)请求，即 QoS 规定具有一个 QoS 等级范围。

(2)可配置的(configurable)，即存在可配置参数集，并能进行自动配置。

针对应用的透明性而言，存在以下两种情况：

(1)网络不能保证应用当前的 QoS 等级时，自适应 QoS 通告应用，由应用用户选择新的 QoS 等级，与网络进行动态重协商，来适应当前网络状态。

(2)应用确定一个 QoS 等级范围，当网络不能保证应用当前的 QoS 等级时，自适应 QoS 自动选择一个新的 QoS 等级，实现自动的重协商，来适应当前的网络状态。它不需要用户干预，具有更好的透明性。

**进一步的研究工作**　我们的研究工作将进一步细化我们提出的自适应 QoS 机制，提出一个自适应 QoS 系统的软件体系结构，一个基于无线环境的实时流应用的自适应 QoS 的结构模型，实现其相应的算法和原型系统。

## 参 考 文 献


1 Davies N, et al. Distributed Systems Support For Adaptive Mobile Applications ACM Mobile Networks and Applications, Special Issue on Mobile Computing - System Services. ACM Press, 1996, 1(4)

2 Blair G S, et al. Quality of service support in a mobile environment: an approach based on tuple spaces. In: Proc. 5th IFIP Int'l Wksp on QoS, 1997

3 Lu S, Lee K-W, Bharghavan V. Adaptive Service in Mobile Computing Environments. In: Proc. 5th IFIP Int'l Wksp on QoS 1997, Chapman & Hall, 1997

4 McIlhagga M, Light A, Wakeman I. Towards a Design Methodology for Adaptive Applications. In: Proc. MOBICOM '98, Dallas, Texas, 1998. 133～144

5 Lu S, Bharghavan V. Adaptive Resource Management Algorithms for Indoor Mobile Computing Environments. ACM SIGCOMM, 1996. 231～242

6 Gecsei J. Adaptation in Distributed Multimedia Systems. IEEE Multi-media, April-June 1997

7 Welling G, Badrinath B. An Architecture for Exporting Environment Awareness to Mobile Computing Applications. IEEE Trans. on Software Engineering, 1998, 24(5): 391～400

8 IETF Network Working Group: RFC 2212 Specification of Guaranteed QoS. S. Shenker, C. Partridge, R. Guerin, IETF, 1997

9 IETF Network Working Group. RFC 1633 Integrated Services in the Internet Architecture: An Overview. R. Braden, D. Clark, S. Shenker, Eds., IETF, 1994

10 IETF Network Working Group. RFC 2475 An Architecture for Differentiated Services. S. Blake et al., Eds., IETF, 1998

11 Katz R H. Adaptation and Mobility in Wireless Information Systems. IEEE Personal Commun., 1st Qtr, 1994, 1(1): 6～17

12 Aurrecoechea C, Campbell A T, Hauw L. A Survey of QoS Architectures. ACM Multimedia Sys. J. - Special Issue on QoS Architecture, May 1998

13 Hutchison D, et al. QoS Management in Distributed Systems in Network and Distributed Systems Management. M. Sloman, Ed., Addison-Wesley, 1994. 273～302

14 Hutchison D, Mauthe A, Yeadon N. Quality of service architecture: Monitoring and Control of Multimedia Communications. Electronics and Commun. Engineering J., 1997, 9(3): 100～106

15 Loyall J P, et al. Specifying and Measuring QoS in Distributed Object Systems. In: Proc. ISORC '98, Kyoto, Japan, 1998

16 Campbell A, Coulson G. A QoS Adaptive Multimedia Transport System: Design, Implementation and Experiences. Media Distrib. Syst. Engineering, 1997, 4: 48～58


(上接第 6 页)


7 Braden B, et al. Resource Reservation Protocol (RSVP) - Version 1 Functional Specification. RFC 2205, Sep. 1997

8 Wroclawski J. The Use of RSVP with Integrated Services. RFC 2210, Sep. 1997

9 Blake S, et al. An Architecture for Differentiated Services. RFC 2475, Dec. 1998

10 Awduche D, et al. Requirements for Traffic Engineering Over MPLS. RFC 2702, Sep. 1999

11 Kompella K, Awduche D O. Notes on Path Computation in Constraint-Based Routing. Internet Draft, draft-kompella-te-pathcomp-00. txt, July 2000, work in progress

12 Jamoussi. Constraint-Based LSP Setup using LDP. Editor, Work in Progress

13 Srisuresh P, Egevang K. Traditional IP Network Address Translator (Traditional NAT). RFC 3022, Jan. 2001

14 Hain T. Architectural Implications of NAT. RFC 2993, Nov. 2000

15 Fuller V, Li T, Yu J, Varadhan K. Classless Inter Domain Routing (CIDR): an Address Assignment and Aggregation Strategy. RFC 1519, Sep. 1993

16 Deering S, Hinden R. Internet Protocol, Version 6 (IPv6) Specification. RFC 2460, Dec. 1998

17 Gilligan, R Nordmark E. Transition Mechanisms for IPv6 Hosts and Routers. RFC 1933, April 1996

18 Carpenter B, Jung C. Transmission of IPv6 over IPv4 Domains without Explicit Tunnels. RFC 2529, March 1999

19 Carpenter B, Moore K. Connection of IPv6 Domains via IPv4 Clouds without Explicit Tunnels. Work in Progress

20 King S, et al. The Case for IPv6. Internet Draft, draft-ietf-iab-case-for-ipv6-06. txt, Dec. 1999

21 Jingguo Ge, Qian Hualin. Cluster-Based Virtual Router, 2001 International Conferences on Info-tech and Info-net Proceedings, Oct. 2001