ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.12, December 2006, pp.2565−2576
DOI: 10.1360/jos172565

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于地址转发表的交换式以太网拓扑发现方法*

孙延涛+, 吴志美, 石志强

(中国科学院 软件研究所 多媒体通信和网络工程研究中心,北京 100080)

A Method of Topology Discovery for Switched Ethernet Based on Address Forwarding Tables

SUN Yan-Tao+, WU Zhi-Mei, SHI Zhi-Qiang

(Multimedia Communication & Network Engineering Research Center, Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: +86-10-51683617, Fax: +86-10-51688152, E-mail: yantao02@ios.cn

Sun YT, Wu ZM, Shi ZQ. A method of topology discovery for switched Ethernet based on address forwarding tables. *Journal of Software*, 2006,17(12):2565−2576. http://www.jos.org.cn/1000-9825/17/2565.htm

Abstract: In this paper, the connections reasoning technique (CRT) based on the predication logic is proposed to infer the connections between network nodes. This technique interprets the address forwarding tables (AFTs) as a set of predicate formulas and translates the topology discovery into a mathematic problem of logic reasoning, so that the topology discovery can be studied by resorting to mathematic tools. An algorithm for topology discovery based on CRT is proposed in this paper. Compared with current discovery algorithms, this method excels in: 1) Applying the redundancies in AFTs more effectively, so that the whole topology can be built up by just small part of AFTs; 2) Naturally resolving the problem of topology discovery for multi-subnet switched domain without any extensions. Furthermore, a method with little cost is proposed to discover the dynamic topology in this paper. This algorithm is successfully applied to the network management system for CBISN (Community Broadband Integrated Services Network).

Key words: network management; topology discovery; switched Ethernets; address forwarding table; topology inference

摘 要: 提出一种称为连接推理技术(connections reasoning technique)的谓词逻辑推理方法推导节点间的连接关系.该方法把交换机地址转发表翻译为一组谓词公式,把拓扑发现问题转变为一个谓词逻辑推理的数学问题,借助数学工具对拓扑发现问题进行研究.基于连接推理技术提出了一种拓扑发现算法,与现有方法相比:(1) 该方法能够更充分地利用不完整地址转发表的冗余信息,只需一小部分转发表就可以把整个网络拓扑构建出来;(2) 该方法完全适用于多子网交换域的拓扑发现.此外,还提出了一种开销很小的动态网络拓扑发现方法.该算法成功地应用在社区宽带综合业务网络管理系统中.

关键词: 网络管理;拓扑发现;交换式以太网;地址转发表;拓扑推理

* Supported by the National Natural Science Foundation of China under Grant Nos.60272078, 60472076 (国家自然科学基金); the Project of the Beijing Committee of Science and Technology of China under Grant No.H020120120530 (北京市科学技术委员会资助项目)

Received 2005-08-23; Accepted 2005-12-14

中图法分类号: TP393 文献标识码: A

交换式以太网(switched Ethernet)是从 5 类双绞线或光纤的集线盒发展起来的一种以太网技术,其组网核心是一个或多个互相连接的以太网交换机(Ethernet switch).每个交换机可以有多个端口,每个端口可与一个交换机或终端设备连接,也可与一个共享式 hub 连接.以太网交换机支持不同端口上的设备并行地进行数据传输,从而大大提高了局域网的传输能力.交换式以太网是目前局域网的主要组网方式之一,其物理拓扑结构发现的目标是确定网络中的各种设备以及这些设备物理端口之间的链路连接关系,因此,物理拓扑发现也称为网络的第 2 层(链路层)拓扑发现.在本文中不加特殊说明,提到的设备间(交换机与交换机之间、交换机与主机之间)的连接关系指的就是设备物理端口之间的连接关系.网络的物理拓扑信息对于网络性能监测与评估、故障发现与定位、资源分配与管理等一系列维护工作具有重要意义.

1 相关工作

研究者已经提出了很多方法[2,3]来发现网络第 3 层,也就是网络层的拓扑结构.3 层网络拓扑不能提供对局域网进行管理及维护所需要的各种拓扑信息.

IETF 于 2000 年推出物理拓扑 MIB(management information base)[4],试图解决网络层以下拓扑结构的发现问题.但是由于没有确定如何获取这些 MIB 对象的机制,关于网络第 2 层(链路层)拓扑结构的自动发现还有待更多的研究.

Myung-Hee Son 等人在文献[12]中提出一种基于生成树协议 STP (spanning tree protocol)的方法,其基本思想是利用生成树的信息构建网桥设备之间的连接关系.该方法的优点在于时间和空间消耗较小,可以发现局域网中的备用链路;缺点是许多交换机不支持生成树协议,适用范围受到一定的限制,并且该方法不能发现交换机和主机之间的连接关系.李涛在更早时候提出了类似的方法[10],但该方法没有对备用链路的发现问题加以讨论.

Richard Black 等人在文献[13]中提出一种基于探测包的方法,其基本思想是,在每个主机上设置一个代理进程,产生一些探测包,并把网卡设置在混杂模式下(在该工作模式下,网卡可以接收到所在网段内的所有数据包),然后,根据每个主机所接收的数据包来判断设备之间的连接关系.该种方法可以判断出交换机之间、交换机和主机之间的连接关系;缺点是在每个主机设备上都要设置一个代理进程,这对一个较大型网络来说是不太可能的事情.

贝尔实验室的 Breitbart 等人提出了基于完整交换机地址转发表 AFT(address forwarding table,简称 AFT)的物理拓扑发现方法[5,6].其算法的核心是直接连接定理:分属两个交换机的一对端口是相连的,当且仅当这两个端口的地址转发条目集合的交集为空,且其并集中包含了该子网中所有交换机的地址条目.该性质的前提条件是每个交换机上的地址转发信息都是完整无缺的.郑海等人[9]提出了一种方法,该方法仅要求只要下行端口的地址转发表是完整的[9],就可以构造出交换机之间的连接关系.直接连接定理不能适用于多子网交换域的拓扑发现[5],2003 年,Bejerano 等人提出一种基于完整地址转发表的多子网拓扑发现算法,很好地解决了此问题[11].该方法首先根据地址转发表和子网信息构造出不同结点之间的粗略路径(skeleton path),然后进一步为每一条路径构造出一组路径约束(path constraint),利用路径约束信息,不断细化粗略路径,最终确定一条唯一的路径.他们证明了该方法的完备性,即如果地址转发表和子网信息可以唯一确定网络拓扑,则使用该方法就可以把这个网络拓扑结构发现出来.

基于完整地址转发表的方法存在一个难以克服的缺点,即需要保证地址转发表的完整性.为了做到这一点,Breitbart[5,6]提出增加额外流量和连续采集两种方法来提高地址转发表的完整性;郑海等人[9]通过在管理站上 Ping 所有交换机的方法来保证下行端口的地址转发表的完整性.在实际网络中,由于地址转发表老化机制的存在、地址转发表长度的限制以及 SNMP 协议采用无连接的 UDP 进行通信,不能提供可靠的数据传输,因此,交换机地址转发表的完整性很难保证,从而影响这些算法的准确性.

在地址转发表中存在着大量冗余信息,如果能够利用这些冗余信息,即使地址转发表不是完整的,也可以把

整个网络拓扑建立起来. Lowekamp 等人在 SIGCOMM 2001 会议上提出了一种基于非完整地址转发表的拓扑发现算法[7].该算法把网络拓扑中的连接分为两种类型:直接连接和间接连接(定义见后文).该算法基于这样一个定理(间接连接定理):如果两个交换机仅有一对端口的通过集相交为空,则这对端口必然相连(间接连接).交换机 S_i 上端口 x 的通过集,是指该网桥上除去 x 以外的其他端口地址转发表的并集.文献[7]提出一个判定两个设备端口是否间接连接的充分必要条件(MKR 定理).但实际上,MKR 定理提出的 3 条规则只是充分非必要条件,其证明过程中犯了一个错误.这一点我们会在后文中详细地加以阐述.

本文试图提出一个完备的规则,如果仅依赖地址转发表,两个交换机之间的端口连接是可以唯一确定的,则利用此规则,就可以唯一确定这两个端口.在此规则的基础上,本文提出了一种有效的算法推导出交换式以太网的物理拓扑关系.该算法要求在拓扑发现期间:(1) 网络拓扑结构是稳定的,设备以及设备之间的连接关系不发生改变;(2) 网络运行状态是正常的,所有设备及链路都能正常工作.

2 拓扑发现算法

2.1 相关概念

为了便于说明算法,下面把本文中用到的一些概念集中介绍一下.

地址转发表:在每个交换机上都维护着一张表,记录着接收的数据包应该从哪个端口转发出去,这张表就是地址转发表.忽略与拓扑发现无关的信息,地址转发表的记录格式可以简单地表示为由端口和 MAC 地址组成的二元组:(*port*,*mac*),其中 *port* 称为转发端口,*mac* 称为转发地址.转发端口为 p 的转发条目构成一个子集,称为端口 p 的转发表.如果交换机端口 p 的地址转发表中包含了该端口所能接收到的所有数据帧的 MAC 地址,则称该端口的地址转发表是完整的.如果一个交换机的每个端口的地址转发表都是完整的,则称该交换机的地址转发表是完整的.地址转发表采用反向学习机制[1]动态建立,并采用超时老化机制把一段时间内没有用到的转发条目删除.

设备类型:在交换式以太网中,我们把网络设备分为 3 类:一类是交换机,每个交换机上维护着一张地址转发表,可以利用 SNMP 读取这张表;一类是主机,在其上面没有地址转发表,通常连接在交换机的某个端口上;还有一类是集线器(hub),这类设备没有 MAC 地址,在网络中不能被直接发现,属于不可管理的透明设备.在局域网拓扑结构中,路由器设备当作一个主机处理.

连接类型:如果两个设备端口不经过其他交换机,直接通过一条物理线路连接在一起,称这两个设备直接连接,或者说这两个设备端口直接连接.如果两个设备端口之间存在着一条路径,我们称这两个设备间接连接,或者说这两个端口间接连接.简单连接的两个设备中间可能会有一个或多个交换机.直接连接是间接连接的一种特殊情况.在下文中,如果不加明确说明,所说的连接都是指间接连接.

管理域和交换域:在进行网络拓扑发现之前,首先要确定管理域(administrative domain),管理域规定了进行拓扑发现的范围.交换域定义为管理域内直接相连的最大的交换机集合,交换域内的交换机必须遵循生成树协议 STP(spanning tree protocol)[1],一个管理域可以包括多个交换域.需要注意的是,一个交换域可能跨越多个 IP 子网.IP 子网定义为任意两个节点不经过路由器通信的最大的 IP 地址集合.不同子网之间的节点通信至少要经过一个路由器.

2.2 理论基础

欲建立起管理域内的整个网络拓扑,可以分为两个步骤:(1) 建立子网之间的连接关系(3 层拓扑).管理域内的各子网通过路由器连接在一起,路由器之间的连接关系可以根据路由器上的路由表信息建立起来;(2) 对每个交换域进行 2 层拓扑发现.第 1 步属于 3 层拓扑发现的范畴,本文的研究工作在第(2)步上.

交换域建模为一棵无向树 $\boldsymbol{G}=(\boldsymbol{D},\boldsymbol{E})$. $\boldsymbol{D}$ 是交换域内所有结点(网络设备)的集合.$\boldsymbol{E}$ 是设备端口之间的直接连接集合,$\boldsymbol{S}$ 表示所有交换机的集合,$\boldsymbol{H}$ 表示所有主机的集合,$\boldsymbol{D}=\boldsymbol{S}+\boldsymbol{H}$.用 $Port(D,p)$ 表示设备 D 的 p 端口,简记为 Dp. $\boldsymbol{E}=\{\langle Ax,By\rangle|A,B\in\boldsymbol{D}$ 且端口 Ax 和 By 直接相连$\}$.

交换机 A 的一个地址转发条目(p,D)可以表示为一个三元组(A,p,D).所有交换机的地址转发表构成一个三元组(A,p,D)的集合.用 $AFT(A,p)$表示端口 $Port(A,p)$的地址转发表中的节点集合.

用四元谓词 $Link(A,x,B,y)$表示设备 A 的 x 端口和设备 B 的 y 端口间接连接,简记为 $AxBy$,其中 A, B, x, y 都是变元.根据地址转发表的含义,从三元组(A,p,D)可以知道交换机 A 经过 p 端口和设备 D 的某个端口 x 间接连接,可以表示为 $Link(A,p,D,x)$.地址转发表的三元组集合可以表示为一组谓词公式的集合 FS (formulas set).如图 1 所示.

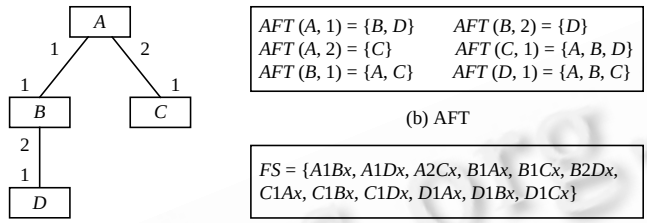


Fig.1 AFTs and formulas set

图 1 转发表和谓词公式集

地址转发表和谓词公式集二者是等价的、一一对应的关系.因此可以抛开地址转发表,从谓词公式集出发,推导设备之间的连接关系.本文提出一种称为连接推理技术 CRT(connections reasoning technique)的谓词逻辑推理方法,推导任意两个节点之间的端口连接关系.

下面给出一组 CRT 方法的基本推理规则 BRR(basic reasoning rule).

(1) 对称律(symmetry rule):$Link(A,i,B,j)\Leftrightarrow Link(B,j,A,i)$.含义是 $Port(A,i)$和 $Port(B,j)$相连当且仅当 $Port(B,j)$和 $Port(A,i)$相连.这一结论显然成立.

(2) 传递律 I (transition rule I):$Link(A,i,C,u)\wedge Link(C,v,B,j)\wedge(u\neq v)\Rightarrow Link(A,i,B,j)$.

证明:由 $Link(A,i,C,u)$可知,从 A 到 C 有一条路径 P_1:$Ai,\ldots,Cu$.由 $Link(C,v,B,j)$可知,从 C 到 B 有一条路径 P_2:$Cv,...,Bj$.又 $u\neq v$,所以 P_1 和 P_2 没有重合部分.因此,从 A 到 B 有一条路径:$Ai,....,Cu,Cv,...,Bj$,可知 $Link(A,i,B,j)$.

□

(3) 传递律 II (transition rule II):$Link(A,i,C,u)\wedge Link(C,v,B,j)\Rightarrow\exists xLink(A,i,B,x)\vee\exists yLink(A,y,B,j)$.其中,$u$ 和 v 是未知数.可能是如图 2 所示的 4 种情况之一(虚线表示间接连接).

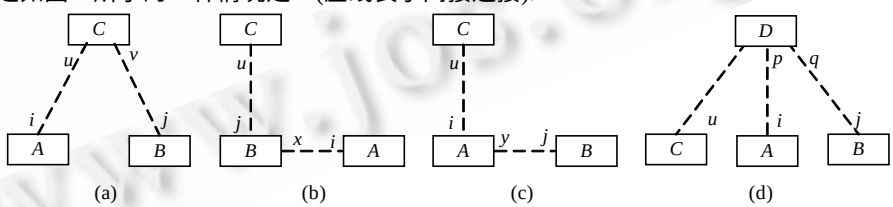


Fig.2 The 4 cases of transition rule II

图 2 传递律 II 的 4 种情况

证明:

a) 如果 $u\neq v$,根据传递律 I 可知,$Link(A,i,B,j)$.

如果 $u=v$,可以分为下面 3 种情况:

b) 从 A 到 C 的路径经过 B,即有一条路径 $Ai,\ldots,Bx,Bj,\ldots,Cu,x\neq j\Rightarrow Link(A,i,B,x)$;

c) 从 B 到 C 的路径经过 A,即有一条路径 $Bj,\ldots,Ay,Ai,\ldots,Cu,y\neq j\Rightarrow Link(A,y,B,j)$;

d) 从 A 到 C 的路径不经过 B,且从 B 到 C 的路径不经过 A,则必然存在一个节点 D,满足 $AiDp\wedge DqBj\wedge p\neq q$,根据传递律 I,可以推出 $Link(A,i,B,j)$.

上面 4 种情况都符合式子的右边. □

(4) 互斥律(exclusion rule):$Link(A,i,B,j)\Rightarrow\neg(\exists x\exists y(Link(A,x,B,y)\wedge(x\neq i\vee y\neq j)))$.

证明:反证法.把结论的否定作为附加前提,假设:$\exists x\exists y((x\neq i\vee y\neq j)\wedge Link(A,x,B,y))$,即在 A,B 之间存在一条路径 $Ax,\ldots,By$.

又 $Link(A,i,B,j)$,可知 A,B 之间存在另一条路径 $Ai,\ldots,Bj$.

因为 $x\neq i$ 或 $y\neq j$,因此,这两条路径不是同一条路径.这与“两个节点之间有且仅有一条路径”矛盾,因此,本结论是正确的. □

(5) 互斥律推论(inference of exclusion rule):$Link(A,i,B,x)\wedge Link(A,y,B,j)\Rightarrow Link(A,i,B,j)$.利用互斥律可以容易地推导出这一结论.

规则(1)~规则(5)构成了 CRT 方法的基本推理规则(BRR).

规则完备性假设(hypothesis for the completeness of BRRs):BRR 规则是完备的,是定义在公式集 FS 上的一组最基本规则,其他推理规则都可以由这些规则推导出来.换句话说,如果一个连接关系可以从公式集 FS 中推导出来,则仅使用 BRR 规则就可以把这个连接关系推导出来.

文献[7]提出称为最小知识需求 MKR(minimum knowledge requirement)的定理.该定理提出一个判定端口连接的完备规则(充分必要条件).在 MKR 中指出:如果两个端口 Au 和 Bv 是间接连接的,当且仅当这两个端口满足至少下面 3 条规则之一:

a) $B\in AFT(A,u)$ and $A\in AFT(B,v)$;

b) $B\in AFT(A,u)$ and $\exists k\neq u$: $AFT(B,v)\cap AFT(A,k)\neq 0$;

c) $\exists i,j,i\neq j$:$AFT(A,u)\cap AFT(B,i)\neq 0\wedge AFT(A,u)\cap AFT(B,j)\neq 0$ 并且$\exists k\neq u$: $AFT(B,v)\cap AFT(A,k)\neq 0$.

应用 BRR 规则,非常容易证明上述规则的充分性,即证明:如果满足上述 3 条规则之一,则可以推导出 Au 和 Bv 是间接连接的.首先把这一问题用谓词公式的形式加以描述(x 表示未知端口):

a) $\{AuBx,BvAx\}\Rightarrow AuBv$;

b) $\{AuBx,BvCx,AkCx\}\Rightarrow AuBv$ ($u\neq k$);或者

c) $\{AuCx,BiCx,AuDx,BjDx,BvEx,AkEx\}\Rightarrow AuBv$ ($i\neq j$, $u\neq k$).

证明:由互斥性推论可知 a)成立.

证明 b):

$BvCx\wedge AkCx\Rightarrow AxBv\vee AkBx$ (传递律 II) ① ①$\wedge AuBx\Rightarrow AuBv$ (互斥律) ②

证明 c):

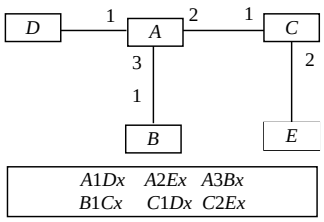
$AuCx\wedge BiCx\Rightarrow AuBx\vee AxBi$ (传递律 II) ① $AuDx\wedge BjDx\Rightarrow AuBx\vee AxBj$ (传递律 II) ②

①$\wedge$②$\Rightarrow AuBx$ (互斥律) ③ $BvEx\wedge AkEx\Rightarrow BvAx\vee BxAk$ (传递律 II) ④

③$\wedge$④$\Rightarrow AuBv$ (互斥律) ⑤ □

上面的证明过程没有使用到传递律 I,这是因为 MKR 规则的必要性存在着问题,即如果两个端口是间接连接的,并不能得出“这两个端口满足至少 3 个条件之一”这一结论.在文献[7]对 MKR 规则必要性的证明中把“两个节点只有一对端口的通过集交集为空”错误地等价于 “两个端口间接连接”.MKR 规则给出的是基于两个节点的地址转发表推导二者连接关系的充分必要条件.该规则不能利用多个(大于 2 个)节点的地址转发表推导连接关系,比如像$\{AuC1,C2Bv\}\Rightarrow AuBv$ 这样使用传递律 I 的推导,在该推导中用到了 A,B 和 C 三个节点的转发表.再举一个复杂的例子(如图 3 所示),要推导 A,B 之间的连接关系,也必须用到多个节点的地址转发表.因此,MKR 规则具有一定的局限性,它不是一种完备的方法.

由上面的讨论中可以看出:(1) MKR 规则仅是 BRR 规则的一个推论,BRR 规则比 MKR 方法具有更为广泛的适用性;(2) BRR 规则可以作为研究网络拓扑问题的一个有效工具.我们认为 BRR 规则是一个完备的规则,其完备性的证明正在进展之中.



Proof:

$A1Dx \wedge C1Dx \Rightarrow A1Cx \vee AxC1$	①
$A2Ex \wedge C2Ex \Rightarrow A2Cx \vee AxC2$	②
①∧②$\Rightarrow A1C2 \vee A2C1$	③
③∧$B1Cx \Rightarrow A1Bx \vee A2Bx \vee AxB1$	④
④∧$A3Bx \Rightarrow A3B1$	⑤

Fig.3 Infering a connection based on multiple nodes' AFT

图 3 基于多个节点的转发表推导连接关系的例子

2.3 连接推导算法

下面介绍利用 BRR 规则进行连接推导的算法.

(1) 算法的数据结构

AFTable 二维表格:*AFTable*[*M*][*N*]用以存储所有交换机的地址转发表,*M*,*N* 分别表示交换机和所有设备的总数.*AFTable*[*i*][*j*]=*k*(*k*>0)对应着地址转发表的一个三元组(*i*,*k*,*j*).*AFTable*[*i*][*j*]=0 表示交换机 *i* 的转发表中没有关于设备 *j* 的表项.*AFTable*[*i*][*i*]=−1,*i*=1,…,*M*.表 1 是一个 *AFTable* 的例子.在表 1 中,*AFTable* 中的 0 忽略不填.

Table 1 The *AFTable* of a topology

表 **1** 某拓扑的 *AFTable*

	A	*B*	*C*	*D*	*E*	*F*
A				2	1	
B	1		2	2	1	3
C	1	1		2	1	1
D	1				1	
E	1			1		
F		1	1			

Connection 连接对象:用于记录一对节点之间的连接关系的对象.比如,*Connection*(*A*,*B*)={*AiBj*}或 *Connection*(*C*,*D*)={*CiDx*∨*CxDj*}.

UndeterminedList(未确定列表):由一个或多个 *Connection* 对象组成的链表,用于记录连接推导过程中尚未确定的 *Connection* 对象.

(2) 连接推导算法的基本思想

首先,根据地址转发表填写 *AFTable*;然后,利用传递律和互斥律计算每一个交换机与其他设备之间的连接关系.

计算一对节点之间连接关系的过程称为计算连接过程,分为 4 步,以计算 *AxBy* 为例:

(i) 使用传递律.借助其他节点利用传递律进行如下推导:{*AuCx*,*CyBv*}⇒*R*1,{*AuDx*,*DyBv*}⇒*R*2,…,{*AuJx*,*JyBv*}⇒*Rn*;

(ii) 使用互斥律:*AxBy*∧*R*1∧*R*2∧…∧*Rn*⇒*Connection*(*A*,*B*).

(iii) 如果 *Connection*(*A*,*B*)比初始值 *AxBy* 更为精确,则进一步利用 *Connection*(*A*,*B*)更新未确定列表中和 *AxBy* 相关的 *Connection* 对象(所谓和 *AxBy* 相关,是指 *Connection* 对象和 *AxBy* 有一个相同端点),比如 *AxCy* 或 *BxCy*.

(iv) 如果上述步骤不能确定 *AxBy*,把 *Connection*(*A*,*B*)加入到未确定列表中.

利用 *Connection*(*A*,*B*)更新未确定的 *Connection* 对象的过程称为更新连接过程,可以分为 3 步,以更新 *Connection*(*A*,*C*)对象为例:

(i) 利用传递律推导 *A*,*C* 之间的关系:*Connection*(*A*,*B*)∧*Connection*(*B*,*C*)⇒*R*1.*Connection*(*B*,*C*) 从 *UndeterminedList* 中得到,或者根据 *AFTable* 得到;

(ii) 利用互斥律更新 $Connection(A,C):R1\wedge Connection(A,C)\Rightarrow R$;

(iii) 如果能够确定 $AxCy$,则把 $Connection(A,C)$从未确定列表中移走;

(iv) 如果 $Connection(A,C)$比初始值更为精确,使用 $Connection(A,C)$继续更新未确定列表.

(3) 补取未确定表项.执行完连接推导算法后,如果 *AFTable* 中还有未被确定的表项,比如 $AFT[i][j]=0$,则需要重新从交换机 i 上读取关于设备 j 的地址转发条目,合并到 *AFTable* 中,并调用“更新连接过程”对其他未确定的连接加以更新.在表 1 中,地址转发信息虽然缺失了很多,但是利用本文的连接推导算法,不需要补取地址转发表就可以把整个拓扑结构(如后文中的图 4,其中 E 和 F 看作交换机)建立起来.

2.4 确定直接连接

利用上节的方法计算出所有节点之间的间接连接关系,就把每个交换机的地址转发表补充完整了.文献[6]指出:如果地址转发表是完整的,就可以确定出设备端口之间的直接连接关系.建立直接连接关系的算法大致可以分为两种,即自底向上的方法和自顶向下的方法.文献[6]中的 *FindLeafConnections* 算法采用的是自底向上的方法,其思路是首先确定网络拓扑树叶节点的连接关系.

自顶向下方法的思想是:首先任意选定一个交换机作为拓扑树的根节点,然后从根出发,逐渐把整个拓扑树建立起来,这种方法一般适合用递归方式实现.该方法把设备端口分为两类:上行端口(根端口)和下行端口.称与根节点相连的活动端口为设备的根端口;其他活动端口为下行端口.每个设备都有一个根端口,主机设备只有根端口,没有下行端口.如果节点 A 连接在节点 B 的下行端口上,则称 A 是 B 的子孙节点,B 是 A 的祖先节点.自顶向下的方法通过确定每个节点最近的祖先节点逐渐逼近与该节点直接连接的祖先节点.自顶向下的方法只需扫描一次整个地址转发表,算法的复杂度和地址转发表的规模 L=交换机的总数 $M\times$所有节点的总数 N 呈线性关系,即为 $O(L)$.算法描述如下:

算法. *BuildTree*(S) /*构造以 S 为根的子树*/

对连接在 S 下行端口上的每个设备(记作 *dev*),作如下操作

(1) 如果 *dev* 的祖先节点没有确定,建立 *dev* 到 S 的连接

(2) 否则

设当前 *dev* 的祖先节点为 A,则 A 和 S 都是 *dev* 的祖先节点.

如果 S 是 A 的子孙节点,建立 *dev* 到 S 的连接.

(3) *BuildTree*(*dev*);

算法结束.

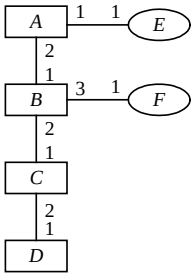
建立起整个网络拓扑以后,如果某个交换机端口上同时连接多个设备,则说明该端口上连接有一个 hub 或哑交换机,拓扑发现算法需要对这一问题作专门处理.

2.5 算法扩展

拓扑发现算法需要处理各种特殊情况,比如网络中存在 hub 设备以及使用 VLAN 技术的情况,对这两种情况的处理请参阅文献[6],本文主要讨论关于多子网交换域和动态网络拓扑的问题.

2.5.1 对多子网交换域的支持

这一问题最先是在文献[5]中提出来的.在单子网交换域中,子网和交换域的范围是一致的,一个完整的地址转发表包含交换域(即子网)内所有节点的地址,因此可以用直接连接定理[6]判定节点之间的连接关系.在多子网交换域中,交换机的地址转发表不能包括交换域内的每个节点地址,比如在图 4 中,一半以上的间接连接关系不能从地址转发表中直接得到.在这种情况下,直接连接定理不再适用,判断直接相连是非常困难的一件事情.比如在图 4 中,交换机 A,D 和路由器 E 构成子网 N_1,B,C 和路由器 F 构成另一个子网 N_2.虽然 $AFT(A,2)\cap AFT(D,1)=\varnothing$,并且 $AFT(A,2)\cup AFT(D,1)=N_1$,但是,$A$ 和 D 并不直接相连.Bejerano 等人摈弃了直接连接定理,使用间接方法解决了这一问题[11].如前所述,需要保证转发表的完整性仍然是该方法难以克服的缺点.本文算法可以在不完整地址转发表上处理多子网交换域的连接问题,其处理过程和单子网交换域没有任何区别.



AFT (*A*, 1) ={*E*}
AFT (*A*, 2) ={*D*}
AFT (*B*, 1) ={*A*, *E*}
AFT (*B*, 2) ={*C*, *D*}
AFT (*B*, 3) ={*F*}
AFT (*C*, 1) ={*A*, *B*, *E*, *F*}
AFT (*C*, 2) ={*D*}
AFT (*D*, 1) ={*A*, *E*}
AFT (*E*, 1) ={*A*, *D*}
AFT (*F*, 1) ={*B*, *C*}

Fig.4 Multi-Subnet switched domain
图 4 多子网交换域

2.5.2 动态网络拓扑

网络拓扑在实际运行过程中会发生一些改变,主要有:(1) 增加节点;(2) 移除节点;(3) 节点移动;(4) 链路状态改变这 4 种情况.拓扑发现算法应该能以较小的代价检测到这些变动.移除一个节点(或节点宕机)比较简单,利用 Ping 操作就可以检测出来.链路状态改变可以根据 SNMP 的 Trap 信息 LinkUp 和 LinkDown[14]或者 MIB 对象 ifAdminStatus 和 ifOperStatus[15]计算出来.通过检测交换机地址转发表的变化,可以定位出新增节点的位置以及检测出节点移动情况.部分节点的变动只会影响到很小一部分地址转发表的变化,因此,也只需要读取每个交换机的一部分转发表,就可以判断出这些变动.为表述方便,用 $F_{A,B}$ 表示节点 *A* 的转发表中关于节点 *B* 的转发条目.

要判断新增节点 *D* 的位置,首先从根节点 *R* 出发,读取转发条目 $F_{R,D}$,确定 *D* 所连接的 *R* 的下行端口 *p*.如果 *p* 上还连接有一个交换机 *S*,则继续读取转发条目 $F_{S,D}$,判断 *S* 和 *D* 的连接关系.这样一直继续下去就可以确定 *D* 的位置.

判断节点移动要复杂一些.我们注意到:如果一个节点发生移动,则在与该节点直连的交换机中,肯定会有一些交换机的转发表中关于此节点的转发条目发生改变.因此,可以通过检查这些变化来发现节点移动.检测并确定节点移位置的算法主要思想是:从根节点出发,遍历每个节点,读取每个节点关于其直连节点的转发条目,判断是否与已知拓扑发生冲突.如果有冲突,比如已知拓扑中 A_i 和 B_j 直连(*A* 是 *B* 的父亲节点),而 $F_{A,B}=(k,B)$,则说明节点 *B* 的位置发生了改变,在实际拓扑中直接或间接连接在 *A* 的 *k*#端口上.判断出某个节点 *B* 发生移动后,通过如下措施定位到此节点的新位置:如果 *B* 节点出现在原父亲节点 *A* 的下行端口上,从节点 *A* 出发,和定位新增节点一样,不断下行,最终定位 *B* 节点的新位置;如果 *B* 节点出现在 *A* 节点的根端口上,先上行,读取 *A* 节点的父亲节点 *C* 关于 *B* 的转发条目 $F_{C,B}$,如此反复,不断上行,直到找到一个节点 *X*,*B* 是 *X* 的子孙节点,然后再从 *X* 出发,和定位新增节点一样,再不断下行,确定 *B* 节点的新位置.

该方法仅采集每个交换机关于其直连节点的转发表条目和变动节点的转发条目,这只是整个转发表的很小一部分,因此开销很小.

2.6 算法比较

表 2 从 3 个方面对本文 CRT 方法和其他基于地址转发表的方法加以比较.首先是比较算法是否要求地址转发表的完整性,这是评判拓扑发现算法优劣的重要方面,因为:(1) 在一个大规模的网络中,地址转发表通常是不完整的;(2) 任何保证地址转发表完整的措施都会极大地增加拓扑发现的时间.其次是算法是否支持多子网交换域,对于一个复杂网络,多子网交换域是不可回避的问题.最后是对动态网络的支持,本文提出了一种开销很小的基于转发表的动态拓扑检测方法.本文算法在这些方面明显优于其他算法.

Table 2 Comparisons of algorithms

表 2 算法比较

	Completeness of AFTs	Multi-Subnet topology	Dynamic topology
CRT	No requirement	Support	Support
Direct connection[6]	Requirement	No support	No support
MKR[7]	No requirement	Partly support	Partly support
Skeleton path[11]	Requirement	Support	No support

3 仿真与应用

3.1 仿真结果

本文采用 NS2 网络仿真软件对不同规模的网络进行仿真,仿真包括 3 个部分:构造网络拓扑、网络流量生成和构造地址转发表.网络拓扑是一个树型结构,根据节点数量(DCOUNT)、交换机的端口数目(12 个)、每个交换机下连接的交换机数目(5 个)和主机数目(7 个)等参数自动生成.网络流量随机产生,其连接时长平均为 5s,按照指数 On/OFF 分布(EXPOO_Traffic)[8]产生数据.节点之间随机地建立 UDP 连接发送数据.节点发送数据的时间占仿真总时间的 20%.仿真完毕后,根据网络中的流量信息采用反向地址学习机制构造出各交换机的地址转发表,转发表的老化时间选取为 120s,180s 和 300s.仿真数据在一台 Pentium IV 2.4GHz、内存为 256M,操作系统为 XP 的计算机进行处理.

在表 3 和表 4 中,DCOUNT/SCOUNT 表示网络中的设备总数和交换机数量;Aging Time 是地址转发表的老化时间;“Item Losts/Ratio”是指转发表条目的缺失数目和比例;“Node Losts”是指转发表的缺失条目涉及到的节点数目;“Before CR”和“After CR”分别表示“连接推理前”和“连接推理后”的统计数据.“Discovery Time”表示拓扑发现时间,分为连接推理时间(CR time)和确定直接连接时间(DC time).表 4 是在 Aging Time=300s 时的统计数据.

Table 3 The statistic of simulations

表 3 仿真结果

DCOUNT/SCOUNT	Aging time (s)	Item losts/Ratio (%)		Node losts		Discovery time	
		Before CR	After CR	Before CR	After CR	CR time (s)	DC time (s)
500/42	120	16000/76.2	154/0.7	500	3	3	
	180	14742/70.2	14/0.0	500	1	1	<1
	300	12776/60.8	0/0.0	500	0	<1	
1000/84	120	71755/85.4	392/0.4	1 000	5	21	
	180	68593/81.7	84/0.1	1 000	1	4	<1
	300	62660/74.6	0/0.0	1 000	0	4	
2000/167	120	304868/91.2	1336/0.4	2 000	8	175	
	180	296000/88.6	334/0.1	2 000	2	65	1
	300	280982/84.1	0/0.0	2 000	0	14	

Table 4 Comparison of simulated results

表 4 仿真结果比较

DCOUNT/SCOUNT	Before CR item losts/Ratio (%)	MKR (after CR)		CRT (after CR)	
		Item losts	CR time (s)	Item losts	CR time (s)
200/17	1217/35.8	1 129	4	0	<1
500/42	12776/60.8	11 749	60	0	<1
1000/84	62660/74.6	58 280	439	0	4

根据表 3 和表 4 的数据,我们可以得出如下几个结论:

(1) 在较大规模的网络中,地址转发表一般是不完整的.这主要是受到地址转发表超时机制和转发表长度等因素的限制.在我们的多次仿真中,转发表的缺失率都在 60%以上.

(2) 与 MKR 方法相比,本文算法的推导能力更强,速度更快.从表 4 中可以看出:本文算法能够推导出更多未知连接,使用的时间更少.比较表 3 中“Before CR”和“After CR”的数据可以看出:经过连接推理后,地址转发表的缺失条目和所涉及的节点数目都有非常显著的下降,仅有极少数未确定连接未被 CRT 推导出来.

(3) 本文算法能够利用少量转发表信息建立起完整的拓扑关系.在 Aging Time=300s 的情况下,地址转发表的丢失率高达 60%~80%,本文算法不需要补取地址转发表就可以把拓扑建立起来.

3.2 应 用

本文算法应用到我们自主开发的社区宽带综合业务网络管理系统(CBISNMS)中,社区宽带综合业务网络(CBISN)是社区内家庭上网、IP 电话和数字有线电视三网融合的解决方案.交换机 SW1200/200/100 提供千兆/百兆线速无阻塞交换,视频组播服务器同时提供 100 个频道电视节目,视频点播服务器同时为 500 个用户提供独立的电视节目,视频会议服务器能够同时提供多个视频会议,家庭网关连接电视机、IP 电话机和计算机,为家庭用户提供电视、电话和计算机上网服务.CBISN 逻辑拓扑如图 5 所示.CBISNMS 网管系统为 CBISN 提供网络管理服务.

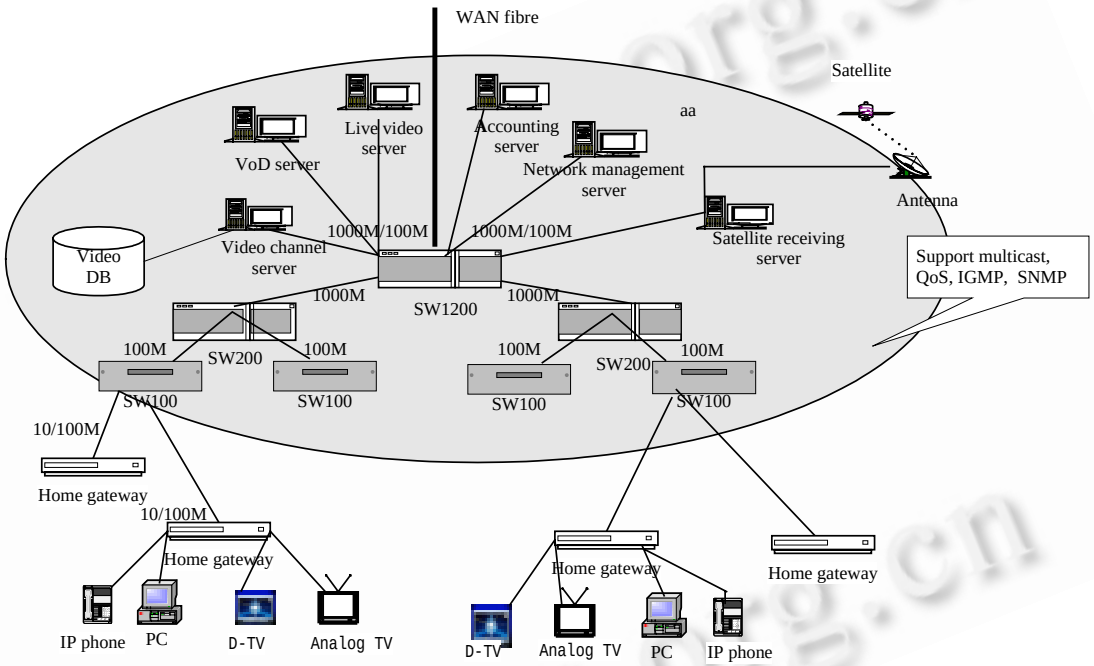


Fig.5 The logical topology of CBISN

图 5 CBISN 逻辑拓扑图

CBISNMS 采用客户机/服务器结构,如图 6 所示.后台服务器运行在 Linux 操作系统上,包括网管后台驻留服务(NM-services)和网管数据库(NM-database),网管数据库采用 PostgreSQL7.3 数据库.前台客户端(GUI)运行在 Windows 平台上,采用 VC++ 6.0 开发.网管后台驻留服务中与拓扑发现相关的模块有:(1) 节点搜索模块(NodeSearcher):搜索管理域内所有的活动节点,判断节点类型;(2) MIB 采集模块(MIBCollector):采集各交换机的地址转发表;(3) 拓扑计算模块(TopoCalculator):根据地址转发表构造出网络拓扑结构.网管前台客户端的拓扑生成器(TopoViewer)负责从网管数据库中读取拓扑信息把网络拓扑图绘制出来.

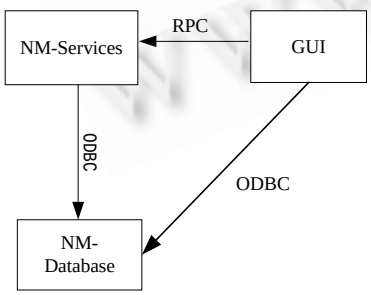


Fig.6 The architecture of CBISNMS

图 6 CBISNMS 构成图

CBISNMS 网管系统拓扑管理部分把网络拓扑分层显示,包括网桥层视图和网桥端口视图,网桥层显示交换机(包括 hub)等网桥设备之间的连接关系,网桥端口视图用来显示交换机各端口所连接的具体设备.如图 7 是网桥层视图,其中:192.168.91.11 是一个千兆交换机;VirtualHUB1406 是一个 hub 设备;其他设备是百兆

交换机,每个交换机(包括 hub)的其他端口上都连接着家庭网关或各种服务器.图 8 所示为是交换机 192.168.91.16 的端口连接情况,选中一个连接(两个设备之间的连线)可以查看所连设备的端口信息.

本文所提算法能够准确、无误地建立起 CBISN 实验网实际物理拓扑结构,图 7 和图 8 是 CBISNMS 网管系统的部分截图.

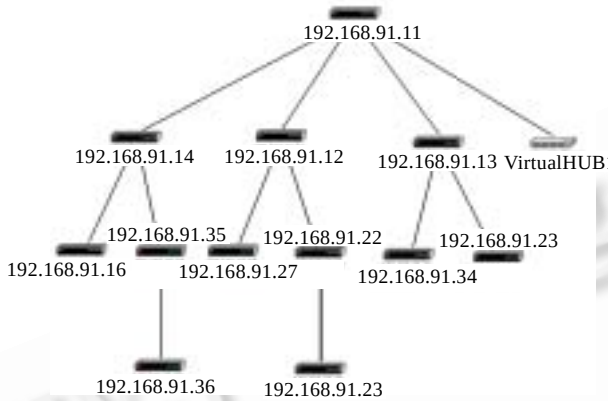


Fig.7 Bridge-Connection view

图 7 交换机连接视图

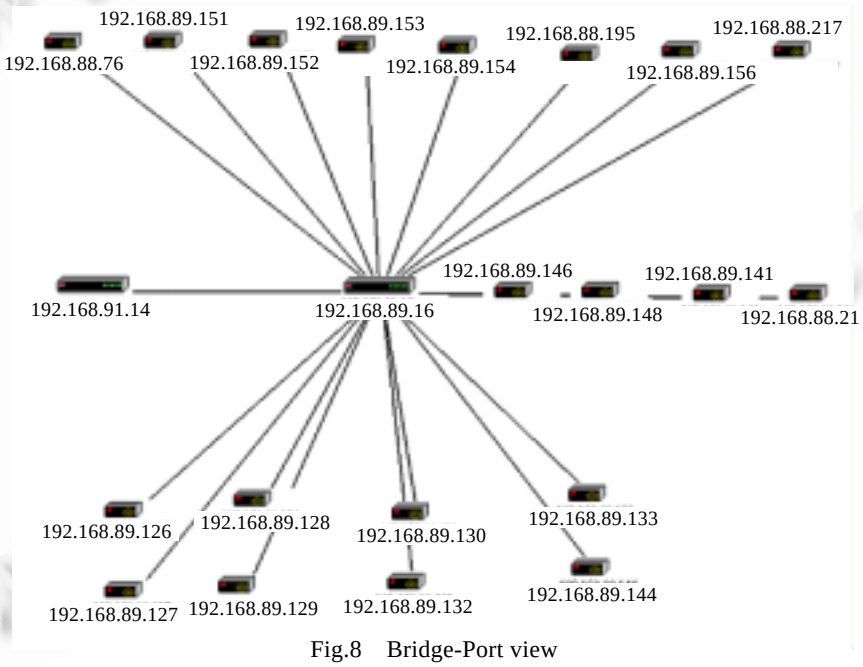


Fig.8 Bridge-Port view

图 8 交换机接口视图

4 结 论

本文提出了一种称为“连接推理技术”的方法,用于发现交换式以太网物理拓扑结构.该方法把地址转发表转化为一组等价的谓词公式,并提出了若干条基本规则用于设备间连接关系的推理.与其他方法相比,本文算法充分利用了地址转发表中的已知信息,把多个交换机上的地址转发条目联合在一起,推导其他缺失的地址转发信息,构造出一个完整的地址转发表.本文算法能够最大限度地减少交换机地址转发表的重新读取次数,从而有

效降低拓扑发现时间.本文算法同样适用于多子网交换域的拓扑发现,其拓扑发现过程和单子网交换域完全一样.本文算法经过简单的扩展就可以支持 VLAN 的拓扑发现,并能够发现 hub 等不可管理的网络设备.此外,在本文中还提出了一种开销很小的动态网络拓扑发现方法.

References:

[1] Tanenbaum AS. Computer Networks. 3rd ed, Beijing: Tsinghua University Press, 1998. 235−236 (in Chinese).

[2] Donnet B, Raoult P, Friedman T, Crovella M. Efficient algorithms for large-scale topology discovery. In: Eager DL, Williamson CL, Borst SC, Lui JCS, eds. Proc. of the ACM SIGMETRICS 2005. New York: ACM Press, 2005. 327−338.

[3] Govindan R, Tangmunarunkit H. Heuristics for Internet map discovery. In: Sidi M, Sengupta B, eds. Proc. of the IEEE INFOCOM 2000. New York: IEEE Press, 2000. 1371−1380.

[4] Bierman A, Jones K. Physical topology MIB. Internet RFC-2922, 2000.

[5] Breitbart Y, Garofalakis M, Martin C, Rastogi R, Seshadri S, Silberschatz A. Topology discovery in heterogeneous IP networks. In: Sidi M, Sengupta B, eds. Proc. of the INFOCOM 2000. New York: IEEE Press, 2000. 265−274.

[6] Breitbart Y, Garofalakis M, Jai B, Martin C, Rastogi R, Silberschatz A. Topology discovery in heterogeneous IP networks: The NetInventory system. IEEE/ACM Trans. on Networking, 2004,12(3):401−414.

[7] Lowekamp B, O'Hallaron DR, Gross TR. Topology discovery for large Ethernet networks. In: Govindan R, ed. Proc. of the ACM SIGCOMM 2001. New York: ACM Press, 2001. 237−248.

[8] Fall K, Varadhan K. The ns manual. http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf

[9] Zheng H, Zhang GQ. An algorithm for physical network topology discovery. Journal of Computer Research and Development, 2002,39(3):264−268 (in Chinese with English abstract).

[10] Li T, Shi ZQ, Wu ZM. Discovering layer-2 topology in bridged local area networks. Computer Science, 2003,30(12):6−8,15 (in Chinese with English abstract).

[11] Bejerano Y, Breitbart Y, Garofalakis M, Rastogi R. Physical topology discovery for large multi-subnet networks. In: Bauer F, Roberts J, Shroff N, eds. Proc. of the IEEE INFOCOM 2003. New York: IEEE Press, 2003. 342−352.

[12] Son MH, Joo BS, Kim BC, Lee JY, Physical topology discovery for metro ethernet networks. ETRI Journal, 2005,27(4):355−366.

[13] Black R, Donnelly A, Fournet C. Ethernet topology discovery without network assistance. In: La Porta T, Ramjee R, Koenig H, Effelsberg W, eds. Proc. of the 12th IEEE Int'l Conf. on Network Protocols (ICNP 2004). Los Alamitos: IEEE Computer Society, 2004. 328−339.

[14] Case J, Fedor M, Schoffstall M, Davin J. A simple network management protocol (SNMP). Internet RFC-1157, 1990.

[15] Case J, McCloghrie K, Rose M, Waldbusser S. SNMPv2. Internet RFC-1905, 1996.

附中文参考文献:

[1] Tanenbaum AS.计算机网络.第 3 版,北京:清华大学出版社,1998.235−236.

[9] 郑海,张国清.物理网络拓扑发现算法的研究.计算机研究与发展,2002,39(3):264−268.

[10] 李涛,石志强,吴志美.桥接局域网第 2 层拓扑结构自动发现.计算机科学,2003,30(12):6−8,15.

孙延涛(1975－),男,山东枣庄人,博士生,主要研究领域为网络管理技术,多媒体通信技术,光网络技术.

吴志美(1942－),男,研究员,博士生导师,CCF 高级会员,主要研究领域为网络和数据通信,多媒体通信.

石志强(1970－),男,博士生,CCF 高级会员,主要研究领域为网络管理,智能光网络,流量工程,仿真,协议工程.