交换式以太网上的多播协议*

王 军+, 吴志美

(中国科学院 软件研究所,北京 100080)

Multicast Protocol over Switch Ethernet

WANG Jun+, WU Zhi-Mei

(Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: 86-10-62645407, E-mail: wyj@isdn.iscas.ac.cn

http://www.iscas.ac.cn

Received 2001-10-16; Accepted 2002-04-10

Wang J, Wu ZM. Multicast protocol over switch Ethernet. *Journal of Software*, 2003,14(3):496~502.

Abstract: Many services, including video conference, whiteboard and video broadcasting, have been running over LANs. However, most of LANs, such as Ethernet, treat multicast just as broadcast, and they all have few supports for multicast. In this paper, a multicast protocol in an LAN switch, named IGMP snooping, is implemented based on VLAN and IGMP. IGMP snooping will be applied to the IP multicast stream control on the switch Ethernet. The basic idea, syntax and semantics of this protocol are given in this paper, and the verification and test procedure is also provided.

Key words: IGMP snooping; multicast; IGMP; CGMP; GMRP

摘 要: 目前,桌面会议、电子白板和视频广播等多播服务大都运行在局域网环境中,而绝大多数局域网结构,如以太网,都采用广播方式处理多播数据,对多播的支持有限.采用 IGMP Snooping 的方法,在二层交换机中设计一个基于 VLAN 和 IGMP 的多播协议,用于控制交换以太网中不断增长的 IP 多播流.描述了该协议的基本思想、语法和语义以及一个该协议验证和测试的过程.

关键词: IGMP snooping;多播;Internet 组管理协议;Cisco 组管理协议;多播组注册协议

中图法分类号: TP393 **文献标识码**: A

目前,在宽带网上的许多新兴业务,例如数据分发、远程学习和分布式数据库等都需要底层网络支持多播通信,在促进有线电视网和计算机网融合的过程中,计算机网也需要提供类似于传统电视网的多点通信能力.现在以太网是最有发展前途的宽带接入方式,在以太网上实现多播通信就成为网络发展的一个必然趋势.传统的以太网由于其通信采用共享介质,多播数据被当作广播来处理,浪费了网络带宽和主机资源.在交换式以太网中,由于引入了 VLAN,可以将通信划分为几个区域,隔离不同的接收者,从而使真正的多播成为可能.实现交换式以太网上的多播与实现 IP 多播相比,只需要实现支持动态组的组管理协议,不需要路由协议.

* Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1998030405 (国家重点基础研究发展规划(973)); the Beijing Committee of Science and Technology under Grand No.H011710010123 (北京市科学技术委员会资助项目)

第一作者简介: 王军(1976－),男,河南汤阴人,博士生,主要研究领域为多媒体数据压缩,网络通信.

Cisco 公司提出了 Cisco 组管理协议 (Cisco group management protocol,简称 CGMP)[1].它是在路由器和交换机之间使用的一种通信协议,主要工作方式是由路由器通过 CGMP 消息通知交换机它所得到的 IGMP 表信息,交换机上的 CGMP 模块会把 IGMP 中的组-主机对转换为组-VLAN 对,并依照这个关系转发多播数据.

在 IEEE 802.1D 中定义了 GMRP(GARP multicast registration protocol),它依赖于 GARP(generic attribute registration protocol)的传输功能,可以用来解决完全二层的多播通信.它是一个纯二层的协议,主机通过 GMRP 向交换机注册多播的 MAC 地址,由交换机维护主机 MAC 和多播 MAC 地址的对应表.

本文提出了一种基于 VLAN 的二层组管理协议,它利用 IP 组管理协议 IGMP 和 VLAN 实现二层和三层多播组的映射,在 3 层组管理协议的基础上实现交换式以太网上的多播.

1 以太网上的多播实现方法

在使用第 3 层多播的情况下,多播接收者使用 IGMP 向本地路由器注册接收 IP 多播数据,路由器使用多播路由协议来构造多播转发树.多播源只传送 IP 多播包的一个副本,路由器接收到数据后,只有在遇到转发树的树叉时,才会再复制一个副本进行传送.由于数据包仅针对每个注册接收者复制,因此这种通信方式其带宽利用效率很高.但是,一旦数据包到达在网络边缘的第 2 层交换机,由于二层交换机无法得到主机与多播组的对应关系,所以它们只能向所有连接的客户机和工作站发送多播数据,使整个 LAN 被广播数据拥塞.

如果需要实现以太网上的多播能力,就需要在以太网交换机中部署新的协议,使它能够得到主机与多播组的对应关系,并将多播组对应到交换端口.

目前被明确规范定义的以太网组播实现有 CGMP 和 GMRP.CGMP 的优点是实现简单,交换机不参与组管理;其缺点是路由器和交换机都需要配置 CGMP.GMRP 的优点是不依赖于路由器,可以和 IP,IPX,AppleTalk 或其他网络层协议协同工作,可扩展性好,转发速度快,支持的组数量多.其缺点也很明显,在使用 GMRP 时,主机的网络接口卡和二层交换机中必须提供对 GMRP 的支持,而现有设备和系统的协议栈对 GMRP 的支持并不充分,并且如果有 3 层多播加入,还必须提供 GMRP 和 IGMP 之间的映射方法.

如果采用 CGMP,路由器和交换机需要部署新的协议,而且必须相互配合一起使用,不利于对现在网络拓扑进行多播扩展.使用 GMRP,目前最大的困难是协议栈并不完善,主机和交换机对 GMRP 的支持不够充分,而且传统多播软件都是在 IGMP 基础上编制的,如果实现二层的多播,还需要实现 IGMP 到 GMRP 的迁移.在不改动原有主机和路由器上的设备和软件的前提下,如何提供二层多播能力,并与原有 3 层多播实现无缝连接是我们要解决的主要问题.

基于上述分析,我们提出使用 IGMP Snooping 的二层多播方法,其基本思想是利用 VLAN 划分多播域,侦听 IGMP 消息用来维护多播 VLAN 表.主要工作方式是:交换机探测以太网包中 IP 包头的协议类型,提取 IGMP 信息包;交换机利用侦听到的 IGMP 信息决定哪一个端口有主机在哪一个组中,为每一个组创建一个 VLAN,创建多播地址和 VLAN 的映射表,并将它添加到 VLAN 表中;交换机向所有的端口(除接收)转发侦听到的 IGMP 查询信息,向属于同一组中的端口转发 IGMP 报告消息,保证路由器和主机之间的 IGMP 协议状态与原来一致;交换机根据 VLAN 表转发多播数据.采用 IGMP Snooping 的好处在于,主机和路由器的软硬件设备不需要进行任何修改,只需在二者之间加一个支持 IGMP Snooping 的交换机即可提供 3 层多播在二层的实现,将多播的范围扩展到边缘子网.表 1 是 3 种实现方法的比较.

Table 1 Comparison of layer-II multicast

表 1 两层组播比较

	CGMP	IGMP snooping	GARP/GMRP
Network layer protocol	IGMP	IGMP	NULL
VLAN	Create/configure VLAN by router's info	Create/configure VLAN by IGMP info	Create/configure VLAN by multicast MAC address
Device	Router and switch	Switch	Switch and host
Response delay	<2s	<1s	<200ms
MAX groups	2^{28}	2^{28}	2^{45}
Multicast address	Mapped from IP multicast address	Mapped from IP multicast address	Defined by GARP PDU
Product	Cisco products	Most of swtich	A few

2 IGMP Snooping

IGMP Snooping 应用的基础是多播 IP 地址与 MAC 地址存在对应关系,以及二层交换机具有 VLAN 功能.VLAN 功能是指交换机能够根据目的 MAC 地址向指定的端口集转发数据.IGMP Snooping 应用的系统协议结构如图 1 所示.

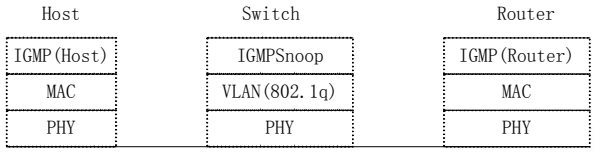


Fig.1 The protocol stack of application system

图1 应用系统的协议栈

IGMP Snooping 的任务是维护 VLAN 与组地址的对应关系,并且能够与多播组的变化同步更新,这样二层交换机就可以按照多播组的拓扑结构转发数据.其功能主要包括侦听 IGMP 报文、维护组地址和 VLAN 的对应表,保持主机 IGMP 协议实体和路由器 IGMP 协议实体的状态一致性, 解决“report flooding”问题.

(1) 侦听 IGMP 信息报文和根据 VLAN 转发数据一般由 Switch 底层提供支持.

(2) IGMP Snooping 分析 IGMP 主机报告,可以得到数据包的源端口和多播地址,然后创建 VLAN,每个 VLAN 包含所有对应于同一个多播地址的端口,从而得到 VLAN 和多播地址的一一对应关系.IGMP Snooping 根据动态组变化和组成员变化更新 VLAN 表和 VLAN 的端口列表,维护 VLAN 和多播地址的一一映射.

(3) 在原有的 IGMP 协议中,当且仅当存在主机处于“成员”状态时,路由器处于“有成员”的状态.这个断言保证了组管理协议的正确性.IGMP Snooping 通过截获主机和路由器的 IGMP 报文,分析其协议状态,然后更新本身数据或状态,产生对应的 IGMP 报文发给对应的协议实体,保证 IGMP 通信双方协议状态转换的正确性,同时实现基于交换机端口的多播.因此 IGMP Snooping 需要保证:

当且仅当存在与端口相连的主机处于“成员”状态时,交换机端口处于“有成员”的状态;

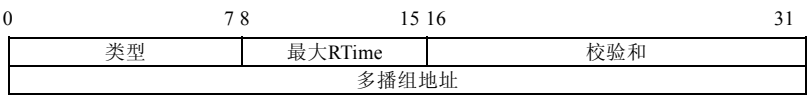
当且仅当存在交换机端口处于“有成员”状态时,路由器处于“有成员”的状态.

(4) 传统 IGMP 协议是针对共享的通信信道,所有的 IGMP 主机协议中都有“报告抑制”的机制,可以解决“report flooding”的问题.即若有多个主机收到查询报告时,多个主机都会启动发送过程(启动随机定时器,等到 Timeout 时再发送报告),先发送的主机报告会抑制其他主机的发送过程(中断其他主机的发送定时器)[1].在交换以太网中,交换机的端口与主机通信使用不同的信道,所以交换机的 IGMP Snooping 实体会收到针对一次查询的多份 IGMP 主机报告.如果都发给路由器(尽管路由器此时只需要一份报告来表明多播组的存在),路由器会认为是针对多份查询出发的报告,会导致协议状态转换出错.另外,大量的主机报告会耗费路由器端口的带宽.因此,IGMP Snooping 在收到另一次 IGMP 查询报文之前,只向路由器转发第 1 个收到的主机报告,其他主机报告只会引起 IGMP Snooping 本身对端口状态的修改,而不会转发给路由器,从而防止在路由器端出现“report flooding”的现象.

3 协议描述

3.1 基本数据类型

(1) IGMP 消息的报文格式[1~3]



类型:

0x11=成员资格查询;0x12=IGMP 版本 1 成员资格报告;0x16=IGMP 版本 2 成员资格报告;0x17=离开组报告.

最大 RTime:规定主机响应成员资格查询前可以等待的最大时间,单位是 0.1s.

校验和:对 8 字节的 IGMP 消息补码之和求 16 位补码.

多播组地址:32 位 IP 组播地址.当“类型”字段为 0x11 时,多播组地址为全 0,表示“常规成员资格查询”报文;当“类型”字段为 0x11 时,多播组地址为 32 位 IP 组播地址,表示“组资格查询”报文,多播组地址是被查询的多播组的地址.当“类型”字段为 0x12 时,多播组地址是被报告的组的 IP 多播地址;当“类型”字段为 0x17 时,多播组地址是要离开的组的 IP 多播地址.

在 IGMP 中,某些组地址具有特定的含义,例如:224.0.0.1(对应于本子网上的所有主机和路由器系统),224.0.0.2(对应于本子网上的所有路由器).

(2) VLAN 表,每个表项格式

VLAN号	端口号	T_{port}

VLAN 号:此 VLAN 在 VLAN 表中的序号(0~4095),16 位.

端口号:在此 VLAN 中的交换机端口号(0~255),8 位.

T_{port}:端口属于此 VLAN 的生存期(单位:s),32 位.

(3) VLAN 与多播组的对应表,每个表项格式

多播地址	VLAN号	T_{group}

多播地址:多播组的地址,32 位.

VLAN 号:与该组相对应的 VLAN 的 VLAN 号,16 位.

T_{group}:多播组的生存期(单位:秒),32 位.

3.2 时间常量与定时器

查询间隔 QI(query interval):125 秒〈为路由器启动查询过程的时间间隔〉.

查询响应间隔 QRI(query response interval):10 秒〈为主机产生报告的最大时间间隔〉.

组成员资格查询 GMI(group membership interval):缺省为 260 秒.

最后组成员查询间隔 LMQI(last member query interval):缺省为 1 秒.

最后组成员查询计数 LMQC(last member query count):缺省为 2.

组计数器 GT(group timer):=LMQI*LMQC,为 2 秒.

T_{port}:用于判断端口的成员资格.收到“常规成员资格查询”报文时启动,此时 $T_{port}=\min(T_{port},R)$,$R$ 等于“常规成员资格查询”报文中“最大 RTime”字段的值*10,在 T_{port} 范围内端口必须收到“成员资格报告”,否则将该端口从多播组中删除.每个 T_{port} 隶属于一个指定组中的一个指定端口(缺省=188s).

T_{group}:用于多播组的生存期,收到“成员资格报告”时启动和更新,如果超时,意味着没有成员希望收到指定多播组的通信,将该多播组删除.每个 T_{group} 隶属于一个指定组(缺省=260s).

3.3 协议的运行机制和过程

为了维护多播组和 VLAN 的对应过程,IGMP Snooping 还使用 4 个通信过程来处理 IGMP 消息和超时事件,通信过程中的主体是路由器(router)、交换机(switch)和 n 台主机(host).这里,假设交换机的每个端口连接一台主机.对交换机端口来说,处理多台主机连接在同一个端口的情况,与处理一台主机的情况完全一致,连接到同一端口的主机之间仍然遵循传统的 IGMP 协议.

(1) 初始化过程

初始时,IGMP Snooping 创建全部主机的组(224.0.0.1,all port)和路由器组(224.0.0.2,与路由器相连端口),设置所有的 T_{port} 和 T_{group} 为无穷大.

(2) 查询和报告过程

在传统的“查询和报告过程”中,路由器会定时发送“常规成员资格查询”,主机收到该消息后会启动本身的计数器,计数器的值为 0 到最大响应时间(报文中的最大 RTime)之间的随机数.在计数器减到 0 时,发送“成员资格报告”.在计数器减到 0 之前,如果收到对应组的“成员资格报告”消息,主机就会停止计数器和发送过程.路由

器在收到“成员资格报告”之后,会重置该组计数器的值,如果在该计数器有效期间,没有收到“成员资格报告”,则该计数器会减到 0,这意味着网络上没有主机在该组中,路由器将不再转发该组的数据.

在 IGMP Snooping 的应用环境中,Switch 在收到路由器发送的“常规成员资格查询”之后,更新 VLAN 表中所有表项中的 $T_{port}=\min(T_{port},R)$,其中 R 是“常规成员资格查询”消息中的最大响应时间(一般为 QRI),然后向“224.0.0.1”组转发该消息.Switch 在收到“成员资格报告”之后,向“224.0.0.2”组转发第 1 个报告,直到下次查询过程开始前,不转发其他报告.这种处理减少了路由器需要处理的报告数,而且不影响路由器对多播组的存在进行正确判断.除了转发第 1 个报文以外,每收到一个 report 报文,IGMP Snooping 都要更新端口的 T_{port} 为缺省值,并且设置 T_{group} 为缺省值.主机和路由器的处理与原有 IGMP Snooping 协议一致.协议流程如图 2 所示.

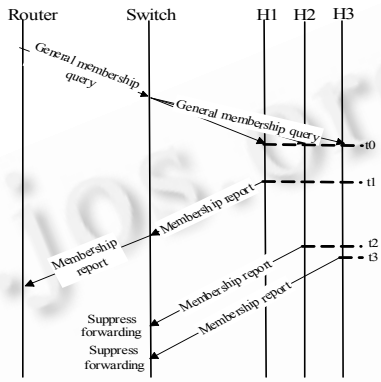


Fig.2 ‘Query Process’of IGMP Snooping

图2 IGMP Snooping的“查询过程”

图中 Router 发送的“常规成员资格查询”被 Switch 转发给每个主机 H_i,H_1,H_2 和 H_3 在 t_0 时刻收到消息并启动定时器来发送“成员资格报告”.H_1 的定时器在 t_1 时刻超时,发送“成员资格报告”,Switch 接收后转发.H_2 和 H_3 分别在 t_2 和 t_3 发送报告,Switch 接收后不会转发.

(3) 离开过程

传统的“离开”过程,希望离开组的主机向路由器发送“离开组报告”,路由器在接收到该消息后,每隔一段时间(=LMQI)发出一条“组资格查询”消息(由 LMQC 决定发出的消息条数),在报文中的最大 RTime 字段中指定了最后响应时间(=LMQI),并且设置组计数器(=LMQI*LMQC),如果没有主机响应查询,组计数器减为 0,路由器将认为没有主机在这个组中,否则在收到主机报告后,路由器会转入正常的查询和报告过程.

在 IGMP Snooping 的应用环境中,Switch 收到“离开组报告”后,向“224.0.0.2”组转发.随后 Switch 会收到路由器发送的“组资格查询”报文,在设置 T_{port}=报文中的最大 RTime 后,向对应组转发该报文.主机和路由器的处理与原有 IGMP Snooping 协议一致.

(4) 超时过程

每过一个时间片,T_{port} 和 T_{group} 都会递减,减为 0 时产生超时消息,在收到主机或路由器的消息时更新 T_{port} 和 T_{group} 的值.

T_{port} 超时的处理:意味着在给定响应时间内,该端口相连的网段没有主机在该组中,因此将对应端口从 VLAN 表中删除.如果对应 VLAN 不包含任何端口,即 VLAN 表中没有该 VLAN 的表项,则表明整个网络没有在该组中的主机,可以将 VLAN 与多播对应表中的 VLAN 和多播表项删除,并停止该组的 T_{group} 定时器.

T_{group} 超时的处理:意味着在给定响应时间内,整个网络没有主机对查询报文进行响应,即该多播组中没有成员,因此从 VLAN 和 VLAN 与多播对应表中删除对应 VLAN 的所有表项.

3.4 状态和状态转换图

主机和路由器上 IGMP 的状态转换图以 IGMPv2 中的描述为准,详细内容参见文献[1,2].

交换机上部署的 IGMP Snooping 实体针对每一个组播组的状态转换图如图 3 所示,其中使用了 256 个 T_{port} 和一个 T_{group} 定时器,有 7 个子过程,触发状态改变的事件有:初始化信号、接收“常规成员资格查询”、接收“组资格查询”、接收“成员资格报告”、接收“离开组报告”、T_{port} 超时和 T_{group} 超时.

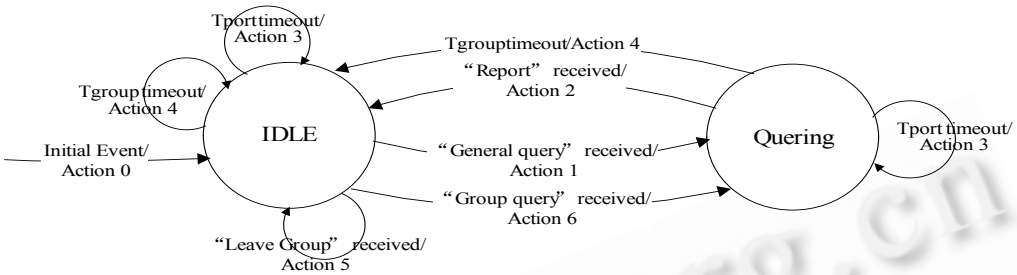


Fig.3 State flowchart of IGMP snooping

图3 IGMP snooping协议的状态流程图

Action 0:添加所有端口到组“224.0.0.1”对应的端口集中,添加路由器所连接的端口到组“224.0.0.2”中,设置所有的 T_{group} 和 T_{port} 为无穷大.

Action 1:向组“224.0.0.1”包含的端口集转发“常规成员资格查询”报文,设置端口的 T_{port} 为 $\min(T_{port},R)$,其中 R 是 IGMP 消息中的最大响应时间.清除所有组的组报告标志.

Action 2:设置报文来源端口在对应组中的 T_{port} 为缺省值.判断是否设置了组报告标志,如果没有,则向组“224.0.0.2”包含的端口集转发“成员资格报告”报文,设置 T_{group} 为缺省值(对应组由报文中的多播组地址标示),设置组报告标志.

Action 3:对应“T_{port} 超时”的处理过程.

Action 4:对应“T_{group} 超时”的处理过程.

Action 5:向组“224.0.0.2”包含的端口集转发“离开组”报文.

Action 6:向对应组转发该“组资格查询”报文,设置对应组的 T_{port}=最后响应时间.

注意:转发是指向除报文来源端口之外的端口发送,因为与该端口相连的主机已经收到了该报文,无须再次发送.

4 验证与测试

因为 IGMP 控制消息也作为组播数据传输,仅依赖二层帧的头信息无法区分 IGMP 和其他组播包,所以交换机必须检查每一个组播数据包以防止漏掉 IGMP 控制消息.如果在低端交换机上使用一个较慢的 CPU 实现 IGMP Snooping,组播数据在高速传输时,对交换机性能和网络状况会带来严重的影响,所以 IGMP Snooping 一般用于带有 ASICs 模块的高速交换机上,采用硬件进行 IGMP 检查.我们为了准确地测试 IGMP Snooping 的可靠性和有效性,在一台二层千兆交换机中设计和实现了 IGMP Snooping 协议,采用 ASIC 芯片进行 IGMP 报文的检查,将所有 IGMP 控制消息从组播数据中提取出来报告给处理器.采用这台二层千兆交换机,我们组建了如图 4 所示的试验网进行实验分析和验证.

频道服务器提供两套节目,一路来自于文件节目源 A,另一路来自于卫星节目源 B.交换机中部署了 IGMP Snooping 协议.测试可以通过侦听网络中数据报和检查设备上的日志,得到动态组变化时数据交换和协议状态变迁的过程.下面是在一段时间内产生的点播过程,表中列出了产生的协议消息、交换机状态和数据交换过程.实验过程中主机 A 连接交换机端口 1,主机 B 连接交换机端口 2,主机 C 连接端口 3,主机 D 连接端口 4,路由器连接端口 15,频道 1 的组播服务地址是 224.5.5.112.在表中对应的 MAC 组播地址用 M112 表示,端口号用数字表示,频道 1 用 C1 表示.

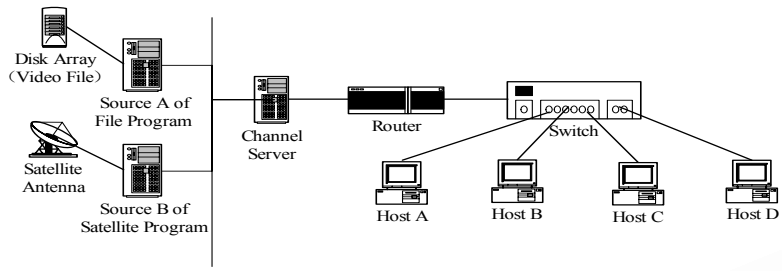


Fig.4 The network topology maps of experiment

图 4 实验环境网络拓扑图

A 点播 C1 之前,与 A 相连的端口 1 不会收到 C1 组播的数据,而在 A 加入 C1 所在组之后,数据被转发到端口 1,同样情况出现在 B,C 和 D.当 B 离开 C1 组后,交换机停止转发 C1 数据到端口 2.这说明 IGMP Snooping 协议可以用在交换机中,配合路由器和主机实现二层的多播功能,实现动态组加入和离开的功能,防止交换机向不在组播组中的网段转发数据.

Table 2 Description of the testing process

表 2 测试过程描述

Demand	IGMP messages	Status of switch	Data flow
Initial	NULL	NULL	NULL
Host A demand channel C1	General membership report (1–>15)	(M112,(15,1))	15–>1
Host B demand channel C1	General membership report (2–>15)	(M112,(15,1,2))	15–>1, 2
Host C demand channel C1	General membership report (3–>15)	(M112,(15,1,2,3))	15–>1,2,3
Host D demand channel C1	General membership report (4–>15)	(M112,(15,1,2,3,4))	15–>1,2,3,4
NULL	General membership query (15–>1,2,3,4) General membership report (1,2,3,4–>15)	(M112,(15,1,2,3,4))	15–>1,2,3,4
Host B leave channel C1	Leave group report (2–>15) Group membership query (15–>2)	(M112,(15,1,2,3))	15–>1,2,3

5 结束语

本文描述了 IGMP Snooping 协议的设计思想和实现方法,并且在一个二层交换机上设计和实现了整个协议,在实验网环境中进行了正确性验证和测试.由上述陈述和实现可见,IGMP Snooping 协议的设计和实现以非常简单、有效的手段解决了目前在交换式以太网中多播应用的网络传输问题,它适合在接入网络交换机中部署,外部通过接入路由器连入 Internet,内部采用高速交换式以太网结构,通过 VLAN 划分虚拟网段.针对智能小区接入中的多播应用 IGMP Snooping 提供了一种有效而可靠的接入网方案,使计算机网络可以很方便地移植有线电视网中提供的一系列广播服务,促进了三网的融合.

当然,在 IGMP Snooping 协议中还存在着一些不足:(1) 如何准确地发现多播路由器所在端口的问题[4],目前采用的方法是通过嗅探路由协议 IP 包来确认路由器,或者在交换机和路由器上部署路由器发现协议,由路由器向交换机报告自己的身份;(2) 考虑在 IGMP Snooping 中加入对 IGMPv3 的支持[5,6],支持基于源的组播;(3) 针对 IGMP 离开组的处理过程,需要优化处理过程,减小离开组的延迟.

References:

[1] Parkhurst WR. Cisco Multicast Routing And Switching. McGraw Hill, 1999. 25~42, 43~53.

[2] Deering S. Host extensions for IP multicasting. RFC 1112, Stanford University, 1989.

[3] Fenner W. Internet group management protocol. Version 2, RFC 2236, Xerox PARC, 1997.

[4] Biswas S, Haberman B, Cain B. IGMP multicast router discovery. Nortel Networks and Cereva Networks. Internet-Draft, 2001.

[5] Cain B, Deering S, Fenner B, Kouvelas I, Thyagarajan A. Internet group management protocol. Version 3, Mirror Image Internet, Cisco Systems, AT&T Labs-Research and Ericsson. Internet-Draft, 2001.

[6] Fenner B, He HX, Haberman B, Sandick H. IGMP-Based multicast forwarding (IGMP proxying). AT&T-Research, Nortel Networks, Internet-Draft, 2001.