维普资讯 http://www.cqvip.com

分布式高性能路由器邻居发现协议实现研究*)

窦睿彧　魏进武　兰巨龙
(国家数字交换系统工程技术研究中心　郑州 450002)

摘　要　邻居发现(ND)协议是网络设备必须支持的协议之一。基于Linux操作系统,本文提出了适合于具有分布式结构的T比特级高性能路由器的一种ND协议实现方案,该方案利用Linux内核提供的netlink机制,分别给出了T比特路由器中先应式地址解析以及主机路由的实现方法,测试结果表明,该方案使得T比特路由器控制平面能够高效可靠地完成邻居发现功能。

关键词　高性能路由器,IP协议栈,IPv6,邻居发现

Research and Implementation on Neighbor Discovery Protocol in Distributed High-performance Router

DOU Rui-Yu　WEI Jin-Wu　LAN Ju-Long
(National Digital Switching System Engineering and Technology Research Center, Zhengzhou 450002)

Abstract　Neighbor Discovery (ND) protocol is one of the protocols must be supported by network device. In this paper a scheme which is fit for high performance Terabyte router with distributed framework is proposed. The scheme is implemented by using netlink mechanism provided by Linux kernel. The implementation of initiative address resolution and host routing are carried out respectively. The results of the test show that the scheme ensured the control plane of Terabyte router performs neighbor discovery function with high efficiency and credibility.

Keywords　High-performance router, IP protocol stack, IPv6, Neighbor discovery

1　引言

IPv6邻居发现协议是IPv4协议族中地址解析协议ARP[2]、ICMP路由器发现协议RDISC(Router DISCovery, RFC 1258)以及由RFC 792中定义的ICMP重定向等协议的综合。路由器和网络主机利用邻居发现协议决定连接在链路上的相邻节点的链路层地址、并察觉链路层地址的变化或消除无效的缓存信息。网络主机和路由器还可利用该协议以判决相邻节点是否可达。网络主机使用邻居发现协议去发现可用于传输数据的相邻路由器。当路由器或到达路由器的路由失效时主机应能重新找到相应的替代节点。由此该协议的主要功能有:路由器发现、前缀发现、参数发现、地址自动配置、地址解析、下一跳确定、邻居不可达检测、重复地址检测以及重定向等。

邻居发现协议使用一系列IPv6控制信息报文(ICMPv6)来实现相邻节点(同一链路上的节点)的交互管理。它使用的五种ICMPv6报文分别是:路由器宣告报文、路由器请求报文、邻居宣告报文、邻居请求报文和重定向。IPv6不再执行地址解析协议(ARP)或反向地址解析协议(RARP),而以邻居发现协议中的相应功能代替,与IPv4地址解析协议比较,IPv6邻居发现协议主要的优点有:

IPv4中的地址解析协议ARP是独立的协议,负责IP地址到链路层地址的转换,对不同的链路层协议要定义不同的ARP协议。IPv6中邻居发现协议(ND)[1]包含了ARP的功能,且运行于因特网控制报文协议ICMPv6上,更具有一般性,且适用于各种链路层协议;

以高效的组播和单播ND报文替代了以往基于广播的地址解析协议ARP、ICMPv4路由器发现和ICMPv4重定向报文;

利用可达性检测是确认相应IP地址代表的主机或路由器是否还能收发报文,对此IPv4没有统一的解决方案。ND中定义了可达性检测过程,保证IP报文不会发送给"黑洞"。

2　分布式高性能路由器中ND协议的功能需求

863重大专项课题"可扩展到T比特的高性能IPv4/v6路由器基础平台及实验系统"的路由器采用共享并行处理器交换式体系结构[5],该结构的基本构成为:线路接口、转发处理、输出处理以及主控和交换结构,其最主要的特点是数据平面与控制平面相分离,即采用分布式转发、集中式处理。结构如图1所示,其中主控是路由器的控制中心,负责运行协议,生成硬件转发数据包所需要的转发表,并下发给转发模块。各个转发模块收到数据包后,硬件实现转发表的线速查找,将其经高速交换网络和输出处理模块送往相应的线路接口模块输出。图2为T比特路由器主控软件总体结构图,在该路由器中需实现双协议栈,本文主要涉及IPv6中的ND功能的实现。

当接口类型为以太网时,为了将数据包正确送到相应的下一跳节点,需要利用ND协议中的地址解析功能来获得下一跳节点的IP地址与链路层地址的映射关系。路由表中所含的是下一跳的IP地址,而生成转发表则需要下一跳的MAC地址,所以也需要利用ND协议中的地址解析功能来获得下一跳节点的IP地址与链路层地址的映射关系。由于线卡中没有支持IPv6的功能,所以ND必须运行在主控上。当一个数据包到达转发板时:如果该包是需要转发的数据包,当

*)基金项目:国家"十五"863计划信息技术领域重大专项(No. 2003AA103510)资助课题。**窦睿彧**　硕士研究生,主要研究方向为高速宽带信息网络技术等;**魏进武**　博士研究生,主要研究方向为Internet网络流量分析及下一代网络体系结构;**兰巨龙**　博士,教授,博士生导师,主要研究方向为高速宽带信息网络技术等。

维普资讯 http://www.cqvip.com

没有查到与之对应的转发表项时就会交由主控处理，直到路由表中下一跳 IP 地址对应的 MAC 地址全被解析出来生成完整的转发表后，查不到转发表的数据包就会被丢弃；如果该包的目的地址和下一跳地址相同即下一跳地址为主机时，转发板将其发往出线卡，若出线卡上没有与此包的目的地址相对应的下一跳 MAC 地址，线卡就会把该数据包发往主控，由于此包将来仍会从出线卡发出去，按照正常情况这样的包会被丢弃从而产生了主机路由问题。前者会使高性能路由器的处理速度变慢，后者会丢弃很多不该丢弃的数据包影响路由器的性能，因此这些问题成为 ND 协议实现中的难点。我们可以通过扩展 ND 协议的地址解析功能来解决上述问题，满足该路由器对 ND 协议的功能需求。在该路由器中，ND 协议的功能需求如下：实现地址解析功能；完成邻居不可达性检测；完成重复地址检测；实现路由器、前缀及参数发现功能；实现重定向功能。

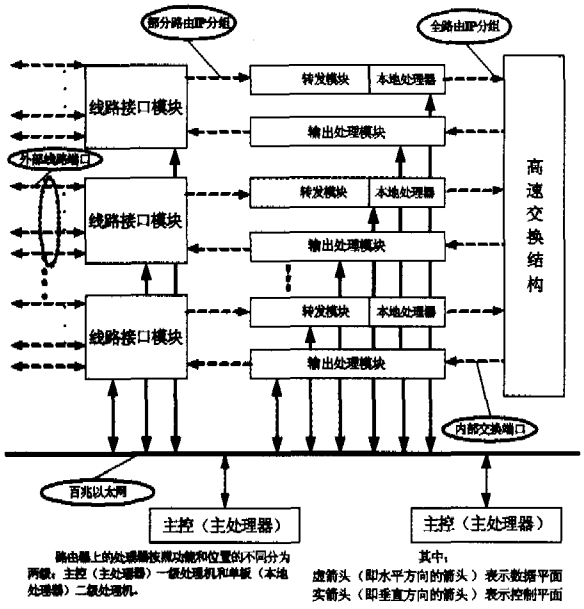

图 1　T 比特路由器总体结构图

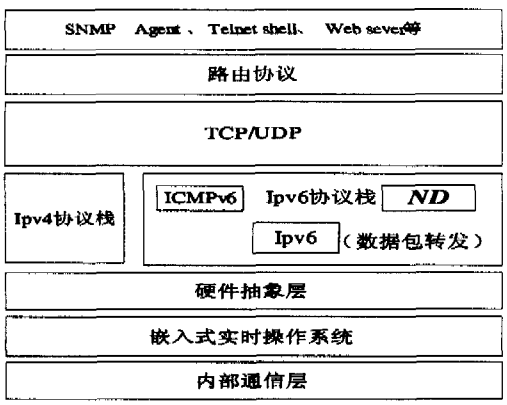

图 2　T 比特路由器主控软件总体结构图

3　ND 协议的实现

3.1　实现方案

针对 ND 的功能需求，我们提出了如下实现方案：主控先把路由表中所有的下一跳 IP 对应的 MAC 地址进行先应式的地址解析，由系统数据维护生成转发表下发给转发板，当需要转发的数据包到来时，该包查到所需表项则直接转发否则就丢弃，这种包不会进入主控，高性能路由器的处理速度得到提高；对下一跳地址为主机的数据包，在虚拟驱动给它置上特殊标志以便在协议栈中进行特殊处理，该处理可以启动地址解析功能并能使该包正常发送而不被丢弃，系统数据维护生成主机映射表下发给线卡，主机路由问题得到解决。

正常情况下，只有当一个数据包需要发送且在内核的邻居表中查不到对应的下一跳 MAC 地址时，ND 协议才会对其进行自动的地址解析。而我们的先应式的地址解析机制是在没有数据包要发送时，对路由表中下一跳 IP 对应的 MAC 地址进行地址解析，我们利用内核中提供的 Netlink[3] 通信机制，将需要解析的路由表中的下一跳 IP 地址，在 Linux 内核中（ND 协议）完成 IP 与 MAC 的对应，并将结果通过 Netlink 机制通告用户进程，并由硬件抽象层完成下发。我们提出的这种 ND 协议的实现方案对 ND 协议进行了扩展实现了先应式的地址解析功能并解决了主机路由问题。该实现方案图如图 3 所示。

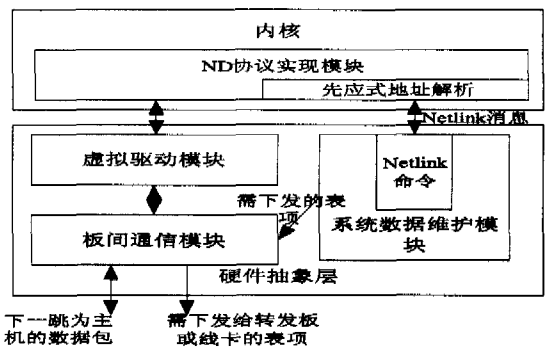

图 3　实现方案图

该方案建立的模型可以描述为：系统数据维护模块对内核空间的输入（关键参数）——下一跳的 IP 地址；虚拟驱动模块对内核空间的输入——下一跳地址为主机的 IP 包；内核空间对系统数据维护模块的输出（关键参数）——与下一跳 IP 地址相对应的 MAC 地址（包括解析出的下一跳地址为主机的 IP 包的 MAC 地址）；内核空间对虚拟驱动模块的输出——得到所需 MAC 地址的下一跳地址为主机的 IP 包。可以看出这里一共有三个流程：一是内核——虚拟驱动模块——板间通信模块——线卡的双向路线，这是下一跳地址为主机的数据包的处理流程；一是内核——系统数据维护模块的双向路线，这是利用内核中提供的 Netlink 通信机制来进行用户空间和内核空间信息交互的流程；最后一个是系统数据维护模块——板间通信模块——转发板或线卡的单向流程，其负责将转发表下发给转发板并将主机映射表下发给线卡。

3.2　方案的分析

方案中，Netlink 的基本原理如下：Netlink 是在内核模块和用户空间之间进行信息交互的进程，它提供了用户和内核空间双向通信的链接。采用 Netlink 机制完成内核与用户空间的双向通信与典型的服务器/客户端模型类似，内核（服务器端）在初始化时创建内核的 Netlink 套接字并监听，用户空间进程（客户端）首先需要创建套接字，经过绑定和连接建立与内核的双向通信。Rtnetlink 则是路由 Netlink 套接字接口，它完成的主要功能包括：建立与读取路由信息、IP 地址与链路参数，邻居的设置、排队规则、流量分类以及数据包过滤器等，这些功能都是通过 Netlink 消息实现的。为了实现 Rtnetlink 的不同功能，可以定义一系列不同的命令，内核在收到不同类型的 Rtnetlink 消息后即可做出相应处理。

维普资讯 http://www.cqvip.com

为满足该路由器对 ND 协议的功能需求，ND 模块中新添加了函数 rtm_getneigh()和 ip6_host_rcv()。ND 模块与硬件抽象层的具体实现通信流图如图 4 所示。

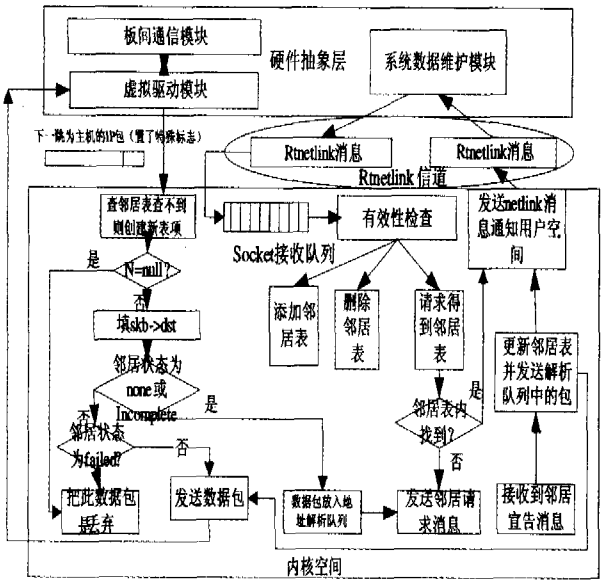

图 4　ND 模块与硬件抽象层的具体实现通信流图

内核从 socket[4] 接收队列（收到的 rtnetlink 消息都放入此队列）中取出 rtnetlink 消息包，先进行有效性检查，然后根据不同的命令（增加、删除或获得邻居表项）得到不同的处理，若是获取邻居表项的命令其对应的操作函数就是 rtm_getneigh()。该函数可根据用户空间的邻居地址解析请求做指定邻居结点的链路层地址的解析。即首先在邻居表中创建一关键字为指定邻居结点 IP 地址的新的邻居表项，同时设置该表项的状态并维护相应的定时器，并根据该地址形成请求结点的地址，构造一仅带 IPv6 基本头的数据包，然后向该邻居结点发送邻居请求消息进行链路层地址的解析。当收到实际存在的邻居结点的邻居宣告消息时，更新相应的邻居表项，并通过 Netlink 报文将解析结果通告给用户空间；否则在发送了 3 次地址解析请求消息仍未收到相应的邻居消息时，将该邻居表项状态置为 NUD_FAILED，并通过 Netlink 报文通知用户空间。

对于下一跳地址为主机的数据包，虚拟驱动在把它送入内核前会置上特殊标志，内核收到这种包就会进行特殊的操作，其操作函数为 ip6_host_rcv()。该函数可以启动地址解析功能且发送此数据包，并把查到的合法表项或解析结果通知用户空间。即先查邻居表，若查不到相应表项就在邻居表中创建一关键字为该数据包对应目的 IP 地址的新的邻居表项，当查到的表项状态为 NUD_NONE 或 NUD_INCOMPLETE 时，发送邻居请求消息，收到邻居宣告消息后，更新内核邻居表把更新后的表项通告用户并发送此数据包；当查到的表项状态为其它合法状态时，把此表项通告用户并发送该数据包；当查到的表项状态为 NUD_FAILED 则丢弃该数据包。

4　功能测试

为了完成该方案的功能测试，现构建如图 5 所示的测试环境。其中，实验机为路由器配置，目标机 1 和 2 为普通主机配置。

4.1　先应式地址解析功能的测试

在如图 5 所示的测试环境下，对先应式的地址解析功能部分进行功能测试时，我们可得到如下的测试结果：当收到用户空间通过 Netlink 信道向内核发送的对某邻居结点的地址解析请求后：

i）当邻居结点被解析地址存在时，ND 模块发送邻居请求消息的次数小于三次就收到邻居宣告消息，然后把此结果通知给了用户进程，内核内邻居表的相应表项也相应更新；

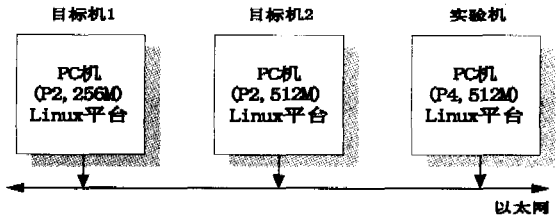

图 5　方案测试环境图

ii）当邻居结点被解析地址不存在时，ND 模块发送邻居请求消息的次数已达三次仍未收到邻居宣告消息，然后通知用户进程解析失败，内核内邻居表的相应表项置为 FAILED。

4.2　主机路由部分功能测试

在如图 5 所示的测试环境下，对主机路由部分的功能测试时，我们可得到如下的测试结果：收到虚拟驱动做了特殊标志的数据包后，内核先查邻居表，若查不到相应表项就在邻居表中创建一关键字为该数据包对应目的 IP 地址的新的邻居表项，当查到的表项状态为 NUD_NONE 或 NUD_INCOMPLETE 时，发送邻居请求消息，收到邻居宣告消息后，更新内核邻居表把更新后的表项通告用户并发送此数据包；当查到的表项状态为其它合法状态时，把此表项通告用户并发送该数据包。

根据 tcpdump 的抓包情况，可以看到：查不到相应表项创建新的邻居表项的情况下，抓到了四个包，分别是：接收到的该数据包、邻居请求包、邻居宣告包、处理完发出的该数据包；查到相应邻居表项时抓到了两个包：接收到的该数据包、查到 MAC 后发出该数据包。

根据这两部分的功能测试结果可以看出：此方案扩展了 ND 协议，实现了先应式的地址解析功能并解决了主机路由问题。

结论　本文以分布式高性能路由器 ND 协议基于 Linux 操作系统的研究为基础，对该协议进行简介并给出了与 IPv4 相关协议比较的优点，重点研究了 ND 协议在高性能 T 比特路由器中的实现，并针对高性能 T 比特路由器中 ND 协议运行在主控时带来的对需要地址解析的数据包处理速度变慢并影响整个路由器处理速度的问题和主机路由问题，提出了 ND 协议在高性能 T 比特路由器中的实现方案并对方案进行了功能测试。

通过对上述测试结果进行分析，我们可得出以下结论：该方案不仅满足了高性能 T 比特路由器对 ND 协议的功能需求，而且实现了内核与用户进程模块的信息交互。从实现的角度来看，实现简单，无需增加额外的模块。

参 考 文 献

1　Narten T, Nordmark E, Simpson W. Neighbor Discovery for IP Version 6 (IPv6)'. RFC 2461, Dec. 1998
2　Plummer D C. An Ethernet Address Resolution Protocol. RFC 826, Nov. 1982
3　Dhandapani G, Sundaresan A. Netlink sockets-overview. The University of Kansas white papers, Sept. 1999
4　李善平，刘文峰，李程远，王焕龙，王伟波. Linux 内核 2.4 版源代码分析大全. 机械工业出版社，2001
5　徐恪，熊勇强，吴建平. 宽带 IP 路由器的体系结构分析. 软件学报，2000(11)