第 33 卷 第 4 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.4 Computer Engineering February 2007
·开发研究与设计技术· 文章编号：1000—3428(2007)04—0283—03 文献标识码：A 中图分类号：TP393.09

基于 CAM 的 10Gbps 线卡地址表维护方案的设计

扈红超，李云涛
(国家数字交换系统工程技术研究中心，郑州 450002)

摘 要：结合国家“863”计划重大专项 T 比特高性能路由器的研制，提出了基于 CAM 的 10Gbps 线路接口卡地址表维护的设计方案，论证了该设计方案的合理性，并给出了线路接口卡地址表维护算法。实验测试的数据结果表明，这种方案满足了 T 比特路由器 10Gbps 线卡的要求。

关键词：T 比特路由器；线卡；CAM

Design of 10Gbps-linecard Address Maintenance Scenario Based on CAM

HU Hongchao, LI Yuntao
(National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002)

【Abstract】This article is based on the work over terabit high performance router of national“863”project, brings up the design of 10Gbps linecard address maintenance scheme, and demonstrates that the scheme is effective for 10Gbps-linecard, finally, it supplies an algorithm of address maintenance of linecard. Test indicates that this meets the requirement of 10Gbps linecard.

【Key words】T-bit router; Linecard; CAM

1 概述

随着 IP 新业务的不断出现、新应用对 QoS 需求的增加，以及光纤到户传输网络的构建，用户对网络带宽需求的不断增加，作为目前 IP 网络核心交换设备的路由器，压力越来越大。解决目前主干网络的压力，一种途径是提高网络的可控性，也就是实施新的网络规划策略来提供更好的服务支持；另一种途径就是扩充主干网络的传输带宽容量。新出现的 10Gbps 接口技术可以满足新的容量需求，同时解决了低带宽接入、高带宽传输的瓶颈问题，扩大了应用范围，并与以前的所有以太网兼容。因此实现路由器 10Gbps 接口，具有重要的意义。路由器的对外接口——10Gbps 线路接口模块结构，如图 1 所示。

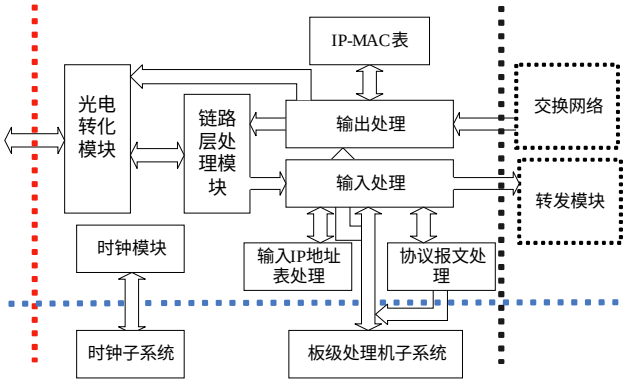

图 1 10G bps线卡模块结构

它主要负责：(1)从外部获得物理信号，提取数据链路层帧并作相应处理；(2)提取 IP 报文并将 IP 包根据类型作相应的处理，将协议报文上交主控，将数据报文送往转发模块；(3)将从调度模块送来的 IP 报文封装成相应数据链路层帧发送到物理链路。根据路由器接口所连接的网络类型，10Gbps 线路接口卡可以分为 10Gbps 局域网(10G-LAN)、10Gbps 广域网(10G-WAN)、10Gbps OC-192C 接口(10G-POS)。作者在 T 比特路由器上实现了这 3 种接口。由于 10G OC-192C 接口的地址表维护比较简单，因此本文主要讨论以太网类型的线路接口。

2 硬件设计分析

10Gbps 线路接口模块主要完成数据链路层和物理层的功能，将输入到线路接口的报文进行分路，一部分数据报文送到转发模块，协议报文上交主控。判断报文是上交还是转发是由 FPGA 查表判断该报文的目的地是不是本路由器决定。同时报文在输出的时候还要在分组头加上数据链路层地址。因此输入、输出模块都有表项存储管理功能模块。按照此功能需求：在 T 比特路由器的设计结构中，10Gbps 以太网线路接口上在输入端有 IP 地址表用来存放本路由器 IP 地址和一些全局地址；在输出端有目的 IP 地址对应的 MAC 地址表用来存放直连机的 IP 地址对应的 MAC 地址。同时由于 T 比特路由器是支持 IPv4/IPv6 的双协议栈路由器。因此，不仅要存放 IPv4 地址，还要存放线路接口的 IPv6 地址。对 IPv6 地址而言，需要由该线路接口识别的地址有单播地址、组播地址和任意播地址，每个线路接口需要 64 条 IP 地址可以满足要求，T 比特路由器支持 512 个线路接口，因此应该存放 512×64＝32 768，再加上 IPv4 的接口地址，共有 65 536(64K)条的容量才能满足要求。在输出端，如果该线路接口所接的二层以太网交换机连接的是以太网子网，按照中等规模的局域网计算，则所需的表项数目也为 64K 个，因此输出查表也

作者简介：扈红超(1983－)，男，硕士生，主研方向：高性能路由器；李云涛，博士生

收稿日期：2006-03-22 **E-mail：**hhc@mail.ndsc.com.cn

选择 64K 的容量。

同时，T 比特路由器要求达到 100MSPS 的查表速度。

2.1 存储 64K 条本机地址要求设计分析

目前流行的查表方案是采用 CAM(Content Addressable Memory)来实现，因此本文总体设计中也采用 CAM 来实现查表处理。作者采用的是 IDT 公司的 CAM(内容寻址存储器)，对于输入查表模块由于只要求 IP 地址的相关操作，因此选择 IDT75K62100 类型的 CAM，其具有 256K*32bit 的存储容量，将所有的 IP 地址按照 IPv6 的地址格式进行存储，可存储的表项条目为：(256K*32bit)/128bit=64K。满足了对表项容量的要求。对输出查表不仅要存储 64K 的 IP 地址表项，而且要存储 64K 的 MAC 地址表项。为了解决这个问题，采用了(CAM+SRAM)的设计方案，就是将 IP 地址存储在 CAM 中，将 MAC 地址存储在 SRAM 中。仍然采用 IDT75K62100 存储 64K IP 地址表项，SRAM 采用 IDT71v658602，其容量为 256K×36 位，其数据总线的接口宽度为 36bit，而 MAC 地址为 48bit 宽，因此采用 36bit*2 来存储一条 MAC 地址，可以存放 MAC 地址的条数为(256K*36bit)/(36bit*2)=128K，因此满足存储 64K 条输出 MAC 地址表项的要求。

2.2 满足 100MSPS 的要求设计分析

为了加快查表速度，将查表系统设计成 FPGA+ IPC+ SRAM 的流水线操作的模式：就是在 FPGA 的控制下，使提出查表请求到输出查表结果分为若干个步骤，一次查表的时间由两个步骤中的最长时间决定。由于 CAM 具有 3 种典型查表请求速度：100M Lookup/s,83M Lookup/s,66M Lookup/s。这里配制成 100M Lookup/s 工作模式。对于存放 MAC 地址的 SRAM 也具有多种工作时钟可以选择，为了和 CAM 模块配合工作，选择其工作的时钟频率为 100MHz，这样在采用流水线工作模式下，一次查表操作过程所需的时间为 10ns，满足了线路接口对地址表 100MSPS 的更新速度的要求。

3 设计流程

由于输入地址表维护方案和输出地址表的维护方案类似，因此本文以输出地址表的维护为例，说明具体的设计流程和算法。

3.1 硬件查表操作的流程

硬件上采用基于 FPGA 的查表方案，通过 FPGA 控制 CAM 来存取数据，同时 CAM 和 SRAM 相连。当单板软件从主控收到一条表项操作请求时，首先通过 MPC860 译码，将命令和数据发送到 FPGA，FPGA 根据下达的操作命令和数据来处理 CAM 中的 IP 地址表项和 SRAM 中的 MAC 地址表项：首先由输出 FPGA 从报头中提取出查表关键字，送给 IPC 查表(搜索)模块，IPC 搜索到与该关键字匹配的表项时，将该匹配表项的地址由 RBUS 送出，通过该地址译码获取 ZBT SRAM 的对应表项，并将查表结果从 ZBT SRAM 的数据总线返回给 FPGA，即得到本次路由查表的结果——数据链层的 MAC 地址。

3.2 软件表维护模块的设计

前面主要讨论 3CAM 硬件接口设计，也就是如何从硬件的角度满足容量和查表速度等方面的要求。事实上，由 FPGA 来对表项进行维护实现难度较大，如果表项更新速度较快，还会影响数据通路的查表操作，造成数据通路的丢包，这是因为硬件本身在维护表项时不能进行流水线操作，尤其是在读 IPC 地址中的内容时只能等上一次的结果返回后才能进行下一次读操作。因此在线路接口设计实现时，表项管理的工作是由板级软件完成的。板级软件作为主控模块和 FPGA 之间的桥梁模块，根据主控模块下发的命令来完成路由器本机地址表的维护(删除或者添加地址表项)工作。为了尽可能减少对硬件的操作，提高硬件操作的效率，在软件模块中安排了一张和硬件一样的 IP-MAC 地址映像表，每条表项由(CAM 偏移地址:IP 地址:MAC 地址)三元组组成，并采用 IP 地址作为查表关键字。因此，对镜像表结构的设计最为关键。在众多的表结构中，哈希表具有查找效率高的特点，其在查找时的平均查找次数为 O(1)，并与所查找的表项数目无关。本文就是基于哈希表来进行大容量表项的管理和维护的。

3.2.1 表项的组织

10G 线路接口卡板级软件输出查表的目的是获得 IP 地址对应的 MAC 地址，因此，将软件里的 IP-MAC 地址表生成 Hash 表项时，由于该表存储大量的表项，因此解决 Hash 表冲突和查表关键字的生成方法影响着 Hash 表的执行效率。Hash 表解决冲突的方法有拉链法和开放地址法，由于拉链法不会产生堆积现象，因此平均查找长度较低；并且拉链法中各个单链表上的结点空间是动态申请的，故它更加适合于造表前无法确定表长的情况，同时单链表添加删除操作比较简单。因此，采用拉链法用来解决冲突问题。

常用的 Hash 表关键字的生成算法有：除余法、异或取关键字法和先异或后除余法。通过在不同的表项容量在 Matlab 下得到了 3 种算法的查表效率，如表 1 所示：其中时间单位为μS，通过比较，可以得出先异或后求余的方法效率最高，因此采用先异或后求余生成关键字的方法设计 Hash 表。

表 1 查表效率对比表

存储容量	大素数	除余法	异或关键字法	先异或后除余法
64K	65 521	1.502 8	1.501 1	1.497 4
128K	131 071	1.253 3	1.248 4	1.248 9
256K	262 139	1.125 0	1.123 6	1.126 0
512K	524 287	1.064 3	1.062 8	1.063 0
1 024k	1 048 573	1.030 5	1.030 4	1.030 6

异或法计算关键字的公式为

Hash[key] = (Addr 32 bit^ Addr 32 bit^ Addr 32 bit^ Addr 32 bit) mod(m)

其中 m 为大素数，在这里取 m 为 524 287。

3.2.2 HASH 表更新算法的设计

为了实现对 CAM 硬件表项的管理，必须在软件中应该记录下 CAM 的存储占用信息。软件在对 CAM 表操作前，首先从该表获得具体的占用信息。所以该占用信息表的存储设计和查找算法影响着内存分配和表项管理效率。在 10Gbps 线路接口卡中，存储设计以位作为最小颗粒度来记录 CAM 空间的占用情况。由于 MPC860 数据线是 32 位宽的，因此，采用字(4B)作为一个存储单元，一个单元可以对应 32 个占用位置，因此需要单元数为(64K/32)=2K，这个记录占用信息表采用一维数组来设计，因此数组容量为 2K＊32bit。为了加快搜索速度，为 32bit 的单元配置了一个位比较数组，该数组是 32*32bit 的结构。数组中的每一项对应 32bit 的某一位，比如 0x00000001 代表了 32bit 的最后一位。查找算法采用轮循和折半算法相结合，这种算法利用数组存储具有连续性和对称性的特点，提高了搜索速度。从后面的试验结果可以看出，这种设计的效率可以满足 10Gbps 线路接口卡本机地址表项更新的要求。该算法的流程为：

(1)Start：初始化数组大小 Range=2016, i =0;

(2)i<Range? 是：goto 3;else 返回 NULL;

(3)搜索占用表的第 1 个字，还有空位置？是：goto 4; else：i++，goto 2；

(4)前两个字节还有空位置？是：goto 5；else：搜索后 2 个字节，goto 5；

(5)第 1 个字节还有空位置？是：goto 6；else：搜索后 1 个字节，goto 6；

(6)找到并返回 end。

这里以删除为例说明地址表的更新过程：板级软件通过板间通信收到主控的删除命令后，从命令中提取要删除的 IP-MAC 地址表项，板级软件根据 IP 地址生成 Hash 关键字，根据 Hash 关键字搜索到该 IP 地址对应的表项，从该表项里可以获得 IP 地址对应的 MAC 地址、该 IP 地址在 CAM 中的索引位置和 MAC 地址在 SRAM 中的位置。根据这些位置信息，板级软件向硬件 FPGA 下达命令，删除对应 IP-MAC 地址表项，整个流程如图 2 所示。

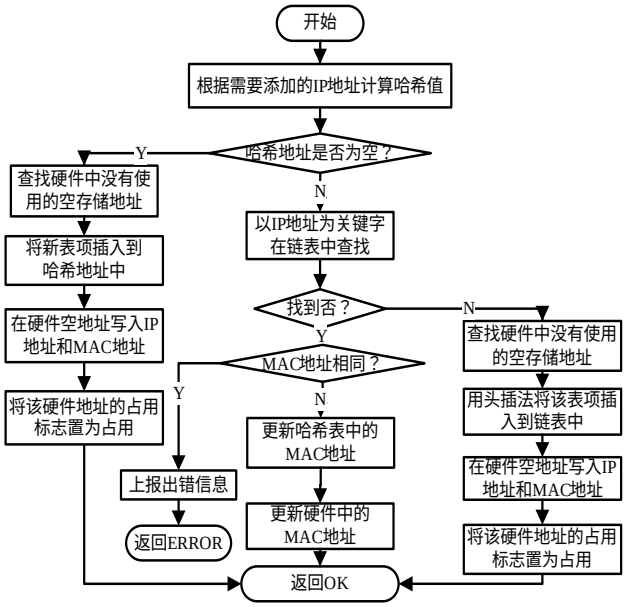

图 2 主控命令处理流程

4 实验测试及结果分析

为了验证算法能不能满足 T 比特路由器的要求，根据 T 比特路由器 10Gbps 测试规范建立了测试环境，在启动单板软件后，由主控模块通过板间通信不断向线路接口模块下达更新地址表命令，得出了下达地址表项条数-时间曲线图，该曲线图描述了更新时间随地址表数目变化之间的关系。从图 3 可以看出，随着表项更新数目的不断增加，表项更新速度越来越慢，这是由 Hash 表本身特点决定的。

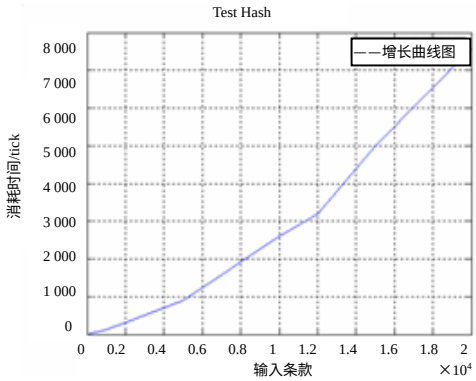

图 3 增长曲线

5 结语

在本文完成的时候，T 比特路由器正在测试中，表项维护功能模块能够达到预期的目标，不过，仅仅利用了 CAM 的写和读这两个基本的操作功能，事实上，IDT 公司生产的 75K62100 类型的 CAM 具有很多扩展功能，可以完成更快速高效的表项更新，如双重写(dual write)、并发写等。充分利用这些功能能够提高维护效率，另外对 CAM 中空位置的搜索算法也可以改进，今后将致力于这方面的工作。

参考文献

1 白建军，卢泽新. 路由器原理与设计[M]. 北京：人民邮电出版社，2002.

2 陈松乔，肖建华. 算法与数据结构[M]. 北京：清华大学出版社，2003.

3 Integrated Device Technology Inc[R]. IDT71V65602 Datasheet. 2002.

4 Integrated Device Technology Inc[R]. IDT72K62100 Datasheet. 2002.

5 WindRiver.Tornado User's Guide(Windows Version)[R]. 1999.

(上接第 282 页)

户界面的机型。该方案如图 4 所示：用一幅预先选好的 I 帧作为图形界面的背景，I 帧上面增加播放的视频小窗口，即 PIG(Picture in Graphic)，使用户在菜单操作时仍能收看到电视节目。菜单内容在菜单层显示，上面覆盖图标层，每个图标可以用单独的调色板，使整个图形界面的色彩丰富美观。

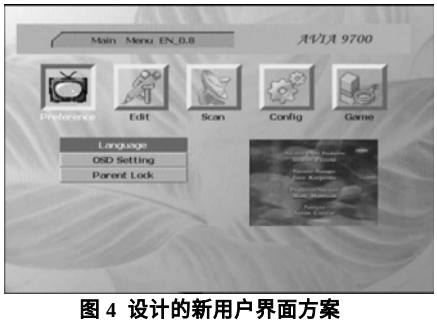

图 4 设计的新用户界面方案

初步测试表明，实验样机的性能稳定。经过计算，资源耗费比一般芯片少，对已有的硬件平台没有提出更高的要求，提高了这款机顶盒的性价比，满足了最初的设计要求，取得了满意的效果。

参考文献

1 章毓晋. 图像处理和分析[M]. 北京：清华大学出版社，1999-02: 17-22.

2 林福宗. 多媒体技术基础[M]. 2 版. 北京：清华大学出版社，2002-09.

3 LSI Logic. SC2005 Single-chip Source Decoder[Z]. 2000-12: 469-474.

4 LSI Logic Chengdu Office. FTA China CWARE Graphics Guide[Z]. 2003-03: 4-8.

5 ETSI EN 302 307 v1.1.1, Draft DVB-S2 Standard[Z]. 2004-06. http://www.dvb.org/documents//en302307.v1.1.1.draft.pdf.

6 LSI Logic. 9600_ERM_3[Z]. 2003-01: 311-322, 345-360.

7 刘政凯，俞能海，张燕翔. 多媒体技术[M]. 合肥：中国科学技术大学出版社, 2001-03: 117-131.