基于双优先级的 IPv4 单播查表解决方法

杨乾斌,张 鹏,陈苏铿,张兴明
YANG Qian-bin,ZHANG Peng,CHEN Su-keng,ZHANG Xing-ming

国家数字交换系统工程技术研究中心,郑州 450002
National Digital Switching System Engineering and Technology Research Center,Zhengzhou 450002,China
E-mail:soldieryang8201@163.com

YANG Qian-bin,ZHANG Peng,CHEN Su-keng,et al.Look-up solution of IPv4 unicast based on double priority.Computer Engineering and Applications,2009,45(5):132-134.

Abstract: In order to deal with the multi-table look-up request caused by IPv4 unicast in large-scale access converging router, combining with the analysis of TCAM+SRAM look-up characteristics,this article brings forward a look-up solution based on double priority,and designs an entry reserved improved select moving algorithm about longest match entry fast update demand. Test shows that this approach can effectively settle the multi-table look-up problem of IPv4 unicast,increase the speed of entry update,economize the FPGA resource and improve the forwarding engine efficiency.

Key words: Ternary Content Addressable Memory(TCAM);look-up table;double priority;entry update;Virtual Private Network(VPN);VPN Routing Forwarding(VRF)

摘 要:针对大规模接入汇聚路由器 IPv4 单播报文的多表查找问题,结合对 TCAM+SRAM 查表技术特点的分析,提出了一种基于双优先级的 IPv4 单播查表解决方法,并对其中的最长匹配表项快速更新需求设计了一种预留表项空间的改进型选择移动算法。测试结果表明该方法能有效地解决 IPv4 单播报文的多表查找难题,提高表项的更新速度,节省 FPGA 资源,提高转发引擎的效率。

关键词:三态内容可寻址存储器;查找表;双优先级;表项更新;虚拟专用网;VPN 路由转发
DOI:10.3778/j.issn.1002-8331.2009.05.038 **文章编号**:1002-8331(2009)05-0132-03 **文献标识码**:A **中图分类号**:TP302.1

随着互联网的发展壮大以及 BGP/MPLS VPN 解决方案的流行,路由器 IPv4 单播报文的查表需求变得愈加复杂。在大规模接入汇聚路由器(large-scale Access Converging Router,ACR)上,由于需要提供对 BGP/MPLS VPN 多种跨域实现方案的支持[1],路由器需要针对 IPv4 单播报文查找 VPN 应用的 VPN 路由转发实例 VRF(VPN Routing Forwarding Instance)表、MPLS 应用的 FTN(Forwarding Equivalence Classes To Next-Hop-Label-Forwarding-Entry)表以及 IPv4 单播表。在报文的转发过程中,查表历来是路由器转发引擎设计的重点,因为查表占用的时间最长,传统的软件式查找算法已经不能适应高速路由器的查找需求,而目前的硬件查找方式可以达到较高的查找速度,但只适合一种类型报文对应单一表项查找需求的场合,IPv4 单播报文多种应用表项的查找需求破坏了这种单一表项查找机制,使得 IPv4 单播报文的处理时序很难与其他类型报文的处理时序对齐。

1 硬件实现的 IPv4 单播报文查表需求分析

目前,大多数路由器都采用了 TCAM+SRAM 的查表结构,这种方式的查表结构将查表关键字存储在 TCAM 当中,而将查表结果存储在 SRAM 当中。当查表模块将查表关键字送入 TCAM 后,TCAM 会返回一个索引地址,将该索引地址送入 SRAM 即可得到查表结果。

1.1 两种典型的 TCAM+SRAM 查表解决方案

应用于路由器 TCAM+SRAM 的查表方式主要有两种[2],分别如图 1 中的(a)和(b)所示。

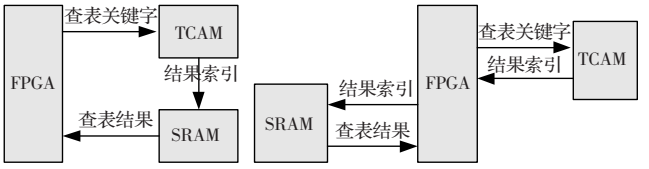


图 1 TCAM+SRAM 的两种组合方式

方式 1 与方式 2 的最大区别在于 TCAM 是否将结果索引送回 FPGA。方式 1 FPGA 的逻辑控制复杂度比方式 2 低,也更节省 FPGA 管脚资源,但最大的缺点是查表段的流水处理时间

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2004AA103130)。
作者简介:杨乾斌(1982-),男,瑶族,通信与信息系统硕士研究生,主要研究方向为宽带信息网络;张鹏(1981-),男,通信与信息系统硕士研究生,主要研究方向为宽带信息网络;陈苏铿(1981-),男,通信与信息系统硕士研究生,主要研究方向为嵌入式处理机;张兴明(1963-),男,教授,硕士生导师,主要研究方向为 T 比特路由器及宽带信息网络。
收稿日期:2008-01-07 **修回日期**:2008-04-08

是 TCAM 搜索表项和 SRAM 根据索引得出查表结果时间的总和，且对 SRAM 的更新与维护操作需要通过 TCAM 进行，会影响到表项的更新速度。方式 2 虽然增加了 FPGA 管脚的使用数量和 FPGA 的逻辑控制复杂度，但查表段可以实现再流水操作，且方式 2 可以由 FPGA 独立对 SRAM 进行更新维护，能够有效地提高表项的更新速度。

1.2　TCAM 查表技术特点

TCAM 的全称是三态内容可寻址存储器(Ternary Content Addressable Memory)。它的特点之一是并行查找技术。并行查找使得 TCAM 的查表时间跟表项的容量无关，无论表项容量多大，只要将查表关键字送入，其得出结果索引的时间是固定的，这也是目前业界广泛采用 FPGA 和 TCAM 技术相结合的一个重要原因。

其次，TCAM 可以实现优先级查找。当 TCAM 搜索到对应关键字的多个匹配表项时，它只将最低地址的匹配表项，也就是最高优先级的索引地址送出，因此当用 TCAM 进行最长匹配查找时，必须将最长前缀匹配的表项放置在 TCAM 的低地址区。另外，TCAM 通过对每条表项提供一个掩码寄存器来实现三态查找，通过掩码寄存器的设置，让 TCAM 在查表时按照约定的格式忽略掉某些数据位，从而只关心特定的比特位。

并行比较使得 TCAM 的功耗比较大，目前，大多数 TCAM 采用 block 的形式组织内部资源，这些 block 可以独立进行工作，且只在工作时才对该 block 供电，以此降低 TCAM 的功耗[3]。以 IDT 的 75K72100 为例[4]，它将整片 TCAM 按其内部组织划分为 32 个 block，每个 block 的容量是 576 K，通过提供的 101 个寄存器来对 TCAM 进行配置。段选择寄存器 SSR(Segment Select Registers)用于按照查表的需求与各种表项容量的要求将 TCAM 进行分段；查表分配寄存器 LAR(Lookup Allocation Registers)用来配置各个 block 的查表位宽；全局掩码寄存器 GMR(Global Mask Registers)则用来配置在查找各个段时屏蔽掉不关心的数据位。

1.3　ACR 中 IPv4 单播查表需求分析

ACR 必须提供对 BGP/MPLS VPN 多种跨域实现方案的支持，同时提供灵活的 VPN 业务配置[1]。对于 IPv4 单播报文，路由器需要首先提取出该报文的入接口号和目的 IPv4 地址查找 VRF 表，若是查表命中，则根据 VRF 表的查找结果转发报文，否则需要再次提取报文的目的 IPv4 地址查找 FTN 表，若 FTN 查表命中则根据 FTN 表的查表结果转发报文，否则将该目的 IPv4 地址作为查表关键字查找 IPv4 单播表，并按照查表结果转发报文。

假设提取查表关键字查找 TCAM 得到结果索引的时间为 t_1，SRAM 根据结果索引得出查表结果的时间为 t_2，则对于上述三种应用的 IPv4 单播报文，报文需要在路由器中缓存的时间分别为 t_1+t_2、$2t_1+t_2$ 和 $3t_1+t_2$，这种报文缓存时间跟报文应用类型有关的查表需求破坏了 TCAM+SRAM 简单固定的处理模式，在单片 TCAM 的情况下只能通过增加报文的缓存时间，加大 FPGA 资源的占用实现与其他类型报文处理时序的对齐。

2　基于双优先级的 IPv4 单播查表设计

TCAM 中的表项在查找时本身就是基于优先级的，因为当有多个匹配表项时，它只将最低地址的索引结果送出，由 1.3 节的分析可知，在 ACR 上，IPv4 单播报文在查找 VRF 表、FTN 表以及 IPv4 单播表时是有优先级的，只有优先级高的转发表查表不中时才能使用优先级较低转发表的查表结果。鉴于此，对于 IPv4 单播报文的多表查找，本文提出了一种基于双优先级的 IPv4 单播查表解决方法，如图 2 所示。

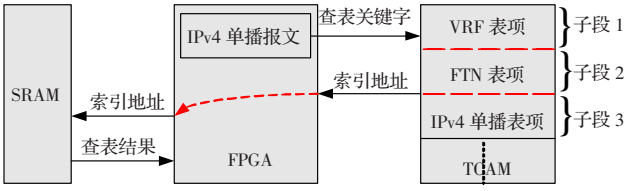


图 2　基于双优先级的 IPv4 单播查表设计

这种查表方法将 IPv4 单播需要查找的 VRF 表项、FTN 表项以及 IPv4 单播表项融合在 TCAM 的同一段(segment)中，在初始化 TCAM 时设置该段的查表位宽为 72 bit，全局掩码寄存器值为与查表关键字入接口号和目的 IPv4 地址相对应的位为 1，其余位为 0，同时根据查找的优先级要求将该段分成 3 个子段，处于最低地址的子段 1 存储优先级最高的 VRF 表项，子段 2 存储优先级次之的 FTN 表项，子段 3 则存储 IPv4 单播表项。对于 VRF 表项，查表关键字为入接口号+目的 IPv4 地址[5]，在向 TCAM 安装 VRF 表项时，将每条 VRF 表项的掩码寄存器设置为与入接口号和目的 IPv4 地址对应的位为有效，其余位无效；对于 FTN 表项和 IPv4 单播表项，查表关键字为目的 IPv4 地址，在安装 FTN 表项和 IPv4 单播表项时，将每条表项对应的掩码寄存器值设置为只与目的 IPv4 地址对应的位为有效，其余位为无效。

当路由器接收到 IPv4 单播报文时，查表模块提取出报文的入接口号以及目的 IPv4 地址送 TCAM 并行查表。如果报文从路由器配置的 VPN 接口进入，则该报文是基于 VPN 应用的，这时 TCAM 会返回查表命中指示和子段 1 中匹配表项的索引地址；若该报文不是基于 VPN 应用的，则在子段 1 不会有匹配信息，因为报文的入接口不会得到匹配，子段 2 和子段 3 中表项掩码寄存器的设置为只有与目的 IPv4 地址对应的位为有效，在搜索子段 2 和子段 3 时查表关键字的入接口号部分已经被掩码屏蔽掉，TCAM 会返回查表命中指示及匹配表项的相应索引地址。

在得到 TCAM 返回的索引地址之后，查表模块将该索引地址送 SRAM 取查表结果，最后向报头处理模块写查表命中标识及查表结果。

由以上分析可知，对于 IPv4 单播报文，不管该报文是基于何种业务应用的，路由器只需提取一次查表关键字查表即可得到查表结果。由于只需要提取一次查表关键字进行查表，在硬件实现上和单一表项的查找机制是一样的，从而避免了 IPv4 单播报文三表优先级查找所带来的过多的缓存资源占用，对齐了 IPv4 单播报文的处理时序，节省了 FPGA 资源。

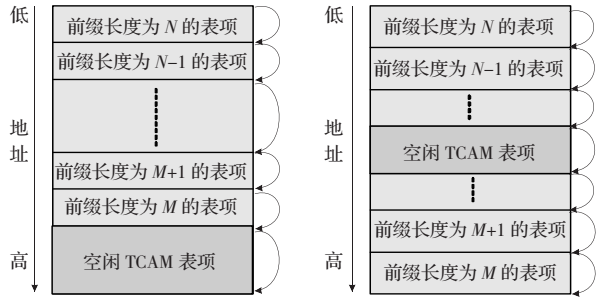
3　基于双优先级的表项更新设计

基于双优先级的 IPv4 单播查表方法存在着很现实的最长匹配查找需求。由于 TCAM 需要将最长前缀的表项保存在前缀较短的表项之前，这种顺序关系使得 TCAM 在对表项进行更新时有可能需要移动其他的表项，造成表项动态更新效率的降低，而在 TCAM 表项更新期间是不能进行查表操作的，报文必

须进入队列缓存来等待查表,因此设计一种快的表项更新算法对 TCAM 进行更新对于节省 FPGA 的缓存资源、提高路由器的性能是显而易见的。

3.1　用于 TCAM 的表项更新算法

对于 TCAM 最长匹配表项的更新,主要有选择移动算法及改进的选择移动算法[6],分别如图 3 中的(a)和(b)所示。



(a)选择移动算法　(b)改进的选择移动算法

图 3　TCAM 最长匹配表项更新算法

TCAM 只要求前缀长度集合块之间的顺序关系,对于每个前缀长度集合块内部各表项之间的顺序关系没有严格要求,选择移动算法就是基于这种思想。假如要插入一条前缀长度为 K ($M\leqslant K\leqslant N$)的表项,则首先从前缀长度为 M 的前缀块开始,将该前缀块的第一条表项移动到空闲 TCAM 区的第一个地址,这样就在前缀长度为 M 的前缀块头部产生了一个空闲地址,于是将前缀长度为 M+1 的前缀块的第一条表项移动到这个空闲地址,以此类推,直到前缀长度为 K 的前缀块产生一个空闲地址为止,将新表项插入到这个地址。

改进的选择移动算法通过将空闲 TCAM 地址放置在 TCAM 的中央来降低算法的复杂度,在最坏情况下,改进的选择移动算法时间复杂度是选择移动算法的一半,算法的工作过程同选择移动算法一样。

令 P_j 表示表项更新时前缀长度为 j(本文只讨论 IPv4 地址的前缀长度,所以 $1\leqslant j\leqslant 32$)的表项更新概率,$\overline{M}$ 表示因表项更新引起的表项移动的平均次数,则改进的选择移动算法的平均移动次数由式(1)给出:

$$\overline{M}=\sum_{j=17}^{32}P_j(j-17)+\sum_{j=1}^{16}P_j(16-j) \tag{1}$$

3.2　预留表项空间的改进型选择移动算法

本文设计了一种预留表项空间的改进型选择移动算法,它通过为每个地址前缀块都预留一些表项空间来提高算法的平均性能,实现原理如图 4 所示。当需要插入一条新的表项时,如果该表项所属的前缀集合块中有空闲地址,则不需要对其他的表项进行移动,将该表项直接插入到这个空闲地址即可;若此时所属前缀集合块中无空闲地址,则需要向上和向下搜索相邻的前缀集合块,直到找到一个最近的空闲地址,这时需要移动一些表项,最后将新表项插入到经过扩展的前缀集合块的空闲地址中。

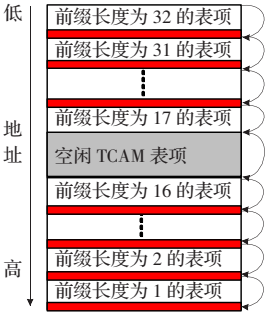


图 4　预留表项空间的改进型选择移动算法

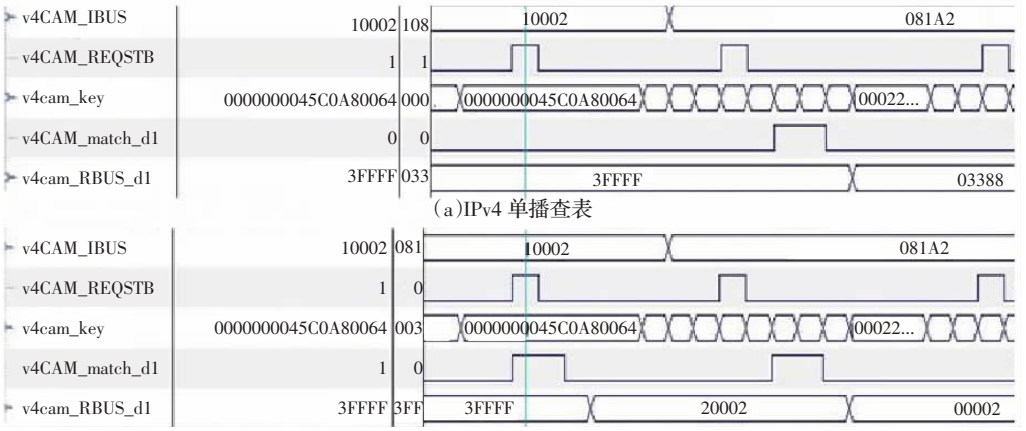
预留表项空间的改进型选择移动算法的平均移动次数由(2)式给出:

$$\overline{M}=\sum_{j=17}^{32}P_j|j-j'|Q(j')H(j')+\sum_{j=1}^{16}P_j|j''-j|Q(j'')H(j'') \tag{2}$$

其中,$Q(i)$表示前缀长度为 i 的前缀块有空闲表项的概率,$H(i)$表示在搜索到前缀长度为 i 的前缀块有空闲表项之前历经的所有前缀块都没有空闲表项的概率。

结合图 4 可知,当 $17\leqslant j\leqslant 32$ 时,j'为 17 到 32 之间的某个值;当 $1\leqslant j\leqslant 16$ 时,j''为 1 到 16 之间的某个值。取极端的情况:即 $j'=17$,$j''=16$,此时意味着除了空闲表项区有空闲地址之外,其余的前缀块区都没有空闲地址,即 $Q(j')=1$、$Q(j'')=1$、$H(j')=1$、$H(j'')=1$,此时式(2)退化成式(1),表明预留表项空间的改进型选择移动算法在最坏情况下的更新复杂度和改进的选择移动算法的复杂度相当。

对于各个前缀块表项空间的预留值,参考 IPMA 的 IPv4 地址前缀分布统计设定[7],以进一步提升算法的平均性能。



(a)IPv4 单播查表

(b)VRF 表查找

图 5　基于双优先级的 IPv4 单播查表解决方法验证

(下转 182 页)