快速路由器的路由查找和流分类算法研究

姚兴苗， 李乐民，胡光岷

(电子科技大学 宽带光纤传输与通信网技术教育部重点实验室 成都 610054)

【摘要】分析了路由器的体系结构发展，研究了路由查找算法和流分类算法在快速路由器中的应用。研究表明，基于分段压缩的路由查找算法支持IPv6路由查找，具有合理的存储容量和快速的查找时间；采用按值分支树的多维综合流分类算法支持前缀和范围匹配，可扩展性强，适合大容量规则数据库。两种算法适合在快速路由器中应用。

关 键 词 体系结构； 路由查找； 流分类； 快速路由器

中图分类号 TP393 文献标识码 A

Research on IP Route Lookup and Packet Classification Algorithms for High Speed Router

Yao Xingmiao，Li Lemin，Hu Guangming

(Key Laboratory of Brodband Optical Fiber Transmission and Communication Networks UEST of China, Ministry of Education Chengdu 610054)

Abstract The development of router architecture is analyzed, and the fast route lookup and packet classification algorithms for high speed router are researched. The research shows the lookup algorithm for IPv6 route lookup with compression trie has reasonable memory space and fast lookup time. The compositive multi-dimensional packet classification algorithm based on tree divided by value is scalable. It can deal with prefixes match and range match for large rule sets. Two algorithms are suitable for high speed router.

Key words router architecture; route lookup; packet classification; high speed router

随着Internet的快速发展和各种宽带技术的不断出现，以及多种Internet业务的增长，路由器的体系结构不断发展，第一代路由器主要采用单处理器共享总线式结构，中央处理器通过通用的总线与多个接口卡互连。中央处理器负责包括路由收集，报文转发处理等所有的事务处理。这种体系结构的性能主要取决于中央处理器的速度和共享总线的带宽，路由器扩展性比较差。第二代路由器在网络接口卡上采用了一些智能处理，如业务接口卡的cache技术来增加转发速率。第三代路由器采用路由与转发相分离的技术，从而有效地解决了路由计算能力的问题，并且总线技术也得到了较大的发展。第四代路由器采用硬件ASIC转发模式和交换结构，解决了带宽容量和性能不足的问题。第五代路由器继承了第四代路由器的优点，增加了更为灵活的网络处理器。对于一些复杂的标准操作，如路由查找算法等，采用硬件协处理器方式提高处理性能，实现软件业务灵活性和高性能硬件转发的有机结合。

路由器技术不断向前发展的同时，也对路由器中的两项关键技术快速路由查找和流分类技术提出新的要求，并且由于传统的IPv4网络需要逐步升级到下一代以IPv6协议为基础的网络，还需要路由查找和流分类对IPv6协议支持，因此研究快速的路由查找和流分类算法在路由器中的应用十分必要。本文从路由器的体系

收稿日期：2004 – 07 – 15

作者简介：姚兴苗(1976 –)，男，博士生，主要从事流分类和路由查找算法方面的研究.

结构发展入手，对快速的IPv4/IPv6路由查找算法和流分类算法进行了研究和讨论，得出了适合在快速路由器采用的快速路由查找和流分类算法。

1 快速路由查找算法

当一个分组到达路由器时，路由器必须根据其目的地址在路由转发表中查找下一跳信息。转发表一般按照如下的形式保存路由项：<目的网络地址/掩码，逻辑端口号>，分组可能匹配多个端口，但分组最终选择所有候选端口中相应掩码最长的端口，这被称为最长前缀匹配(能够有效的降低路由表的大小，并且在一定程度上缓和IPv4地址的枯竭问题)。寻找高效的路由表查找算法是相当困难，查找算法的性能不仅要考虑到快速查找时间，还要求低存储空间和快速路由表更新。路由查找算法大致分为3类：

1) 基于三态内容可寻址存储器(Ternary Content Addressable Memory, TCAM)的算法：用TCAM来实现路由查找非常简单，只需要一次查找。如果存在多个匹配表项，在经过优先级比较之后，就可确定下一跳信息。目前的TCAM存储器容量较小，而且价格昂贵，另外，TCAM部件的功耗很大，不利于硬件集成。在克服TCAM的缺陷后，采用TCAM实现路由查找是一个不错的方案。文献[1,2]提出了几种不同的基于TCAM的查找算法。

2) 基于Hash的算法：文献[3]提出了基于地址长度的二分Hash查找算法，针对地址最大长度为W比特的路由查找，需要$O(\mathrm{lb}W)$次哈希表查找，需要的最大存储空间为$O(NW)$，其中N为路由表中的表项数目。但是算法不易硬件实现，并且如何寻找高效的哈希函数还需进行研究。

3) 基于Trie的算法[4]：Trie又称为数字查找树(DST)。最简单的Trie就是二叉查找树。查找树算法通常可以利用软件实现。但它的查找速度慢，最坏情况下，所需要的查找次数是$O(W)$次， 二叉树的数据结构需要的存储空间在最坏情况下为$O(NW)$。路径压缩树和级压缩树对二叉查找树作了改进，但是其动态性能较差。

基于多比特查找树(Multibit Tries)的算法一次查找多个比特[5,6]，与基本的二叉树相比大大降低了查找时间复杂度。多比特查找树的分段式硬件查找算法将多比特查找树中的节点通过简单的地址映射与路由表项的硬件存储地址关联，从而达到高速的查找速率。典型的查找算法是文献[7]提出的DIR 24-8方案。分段式硬件查找算法具有查表速率快，结构简单，易于实现，对硬件要求不高等优点，适合在快速路由器上实现。但多数算法专门针对IPv4路由查找，不易扩展到IPv6查找。文献[8]提出的算法是一种基于分段路由查找方法，可扩展到IPv6查找。但是它针对IPv6查找所需存储空间是令人难以接受的。文献[9]提出了一种基于分段压缩的方案，与文献[8]提出的算法相比，该算法能进行空间压缩，算法需要存储空间小，并且算法的平均查找时间少。因此，算法能够在需要对IPv6路由查找支持的快速路由器上实现。表1所示列出了几种路由查找算法的性能。

表1 几种路由查找算法比较

算法	查找复杂度	更新复杂度	存储空间	备注
TCAM方案	$O(1)$	$O(N)$	$O(N)$	
二分Hash查找	$O(W)$	$O(\mathrm{lb}W)$	$O(W\mathrm{lb}N)$	
二叉Tire树	$O(W)$	$O(W)$	$O(NW)$	
多比特Tire树	$O(W)$	$O(W/K+2^K)$	$O(2^KNW/K)$	
DIR 24-8方案	$O(2)$	$O(2^{16})$	33 MB	
LLCAT方案	$O(W/K)$	$O(W/K+2^{(K-1)})$	$O(2^KNW/K)$	支持IPv6
分段压缩方案	$O(W/K)$	$O(W/K+2^K+M)$	$O(2^KNW/KM)$	支持IPv6

2 快速流分类查找算法

随着路由器的发展和多种业务的需求，如不同的用户需要不同的Qos要求，一些如接纳控制，资源预留，公平调度等新技术需要在路由器上实现。而实现这些技术的前提条件是需要对分组进行分类。分类主要依

据分组中第三层分组头中的第四层协议类型、源地址和目的地址，第四层报文头中的源端口和目的端口号等信息，有时候还包括第二层的MAC地址和更上层的头信息乃至分组的内容。流分类算法大致可以分为以下3类：

1) 基于硬件的查找算法：基于硬件TCAM的流分类算法[10, 11]，算法性能要受到TCAM缺陷的制约。重复流分类算法是一种多维的流分类算法[12]，其优点是查找时间快并能同时支持前缀和范围匹配，但是需要的存储空间太大，算法不易扩展。位向量算法便于硬件实现[13]，在小规则数量时具有快速的查找速度，但查找时间随着规则数量的增加线性增加。聚合位向量算法使用聚合位向量的方法改进了BV算法的查找时间[14]，但是所需要的存储容量增大。

2) 基于查找数的算法：这一类流分类算法的主要特征是：在预处理时建立以查找树为中心的数据结构，流分类时通过一次或多次访问查找树得到分类结果。这类方法如查找树网格等[15]。该类算法的空间复杂性相对较小，但算法数据结构较为复杂，不易硬件实现。

3) 基于Hash的算法：流分类的Hash表方法有2种实现途径，(1)是直接使用Hash函数实现流分类。(2)是预先对流分类规则作某些处理，以达到降低冲突率的目的，如文献[16]提出的元组空间搜索算法。但还需研究针对流分类有效的Hash函数。

由于流分类规则数量和规则维数不断扩展，上述的流分类算法不能完全满足路由器要求。为此流分类算法应该综合使用多种方法，按流分类过程本身的特点，将其划分为多个阶段，根据每个阶段特点和流分类总体目标进行不同处理。Modular算法实际上就是这样的一种算法[17]，它将查找分为三个阶段，对每一阶段使用不同的优化算法。Modular算法将每一维规则看作前缀的形式，使它能够适应规则的扩展。然而它只支持前缀匹配，不能够直接支持范围匹配。在多维的规则匹配中，有些维是需要范围匹配的。同时由于Modular方法使用索引跳转表，某些规则可能在子树和叶子规则束中重复多次，可能会引起存储空间的爆炸。针对Modular算法存在的问题，文献[18]提出了一种适用于多维、大容量、可扩展的按值分支树高效流分类查找算法，算法同时支持前缀匹配和范围匹配，能处理大型的流分类数据库。由于对于规则每维的匹配类型没有严格的要求，规则的维数也容易扩展。算法不仅可以软件实现，还可实现硬件的查找，是一种能在快速路由器上应用的流分类算法。表2所示为几种流分类算法比较。

表2　几种流分类算法比较

算法	支持的规则类型	支持的元组数	算法类型	备注
TCAM方案	前缀匹配	多元组	硬件算法	硬件CAM
重复流分类算法	前缀、范围匹配	多元组	可硬件实现	
位向量	前缀匹配	多元组	硬件算法	
查找树网格	前缀匹配	二元组	软件算法	
元组空间搜索	前缀匹配	多元组	可硬件实现	
Modular算法	前缀匹配	多元组	可硬件实现	
按值分支树算法	前缀、范围匹配	多元组	可硬件实现	

3 结 束 语

本文对快速路由查找和流分类算法进行了讨论，并指出了快速路由器对路由查找和流分类算法的要求。随着路由表数量和流分类规则数量的增多，基于分段压缩的快速路由查找算法和综合的流分类算法更适合在快速路由器中应用。另外，考虑要支持IPv6协议，路由查找算法和流分类算法还需具有对协议的扩展性要求。

参 考 文 献

[1] Ravikumar V C, Mahapatra R N. TCAM architecture for IP lookup using prefix properties[J]. IEEE Micro, 2004, 24 (2): 60-69

[2] Liang Zhiyong, Wu Jianping, Xu Ke. A TCAM-based IP lookup scheme for multi-nexthop routing[C]. International Conference on Computer Networks and Mobile Computing, Shanghai, China, 2003. 128-135

[3] Waldvogel M, Varghese G, Turner J, *et al*. Scalable high speed IP routing lookups[C]. Proceedings of ACM SIGCOMM ’97, French Riviera, 1997. 25-36

[4] Tzeng HH-Y, Przygienda T. On fast address-lookup algorithms[J]. IEEE Journal of Select Areas in Communication, 1999, 17(6): 1 067-1 082

[5] Sahni S, Kun Suk Kim. Efficient construction of multibit tries for IP lookup[J]. IEEE/ACM Transactions on Networking, 2003, 11(4): 650-662

[6] Jia Jinpeng, Lin Chuang, Liu Weidong. A fast two-way IP lookup algorithm based multibit-trie[C]. International Conference on Computer Networks and Mobile Computing, Shanghai, China, 2003. 136-142

[7] Gupta P, Lin S, McKeown N. Routing lookups in hardware at memory access speeds[C]. Proceedings. IEEE INFOCOM '98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies, San Francisco 1998. 1 240-1 247

[8] Yilmaz P, Belekiy A, Uzun N, *et al*. A Trie-based algorithm for IP lookup problem[C]. IEEE GLOBECOM2000, San Francisco, 2000. 593-598

[9] Yao Xingmiao, Li Lemin, Hu Guangming. A fast IPv6 route lookup algorithm with hash compression[C]. 2004 International Conference on Communications, Circuits and Systems, Chengdu, China, 2004. 674-677

[10] Van L J, Engbersen T. Fast and scalable packet classification[J]. IEEE J. on SAC, 2003, 21(4): 560-571

[11] Liu Huan. Efficient mapping of range classifier into ternary-CAM[C]. Proceedings of 10th symposium on High Performance Interconnects, California, 2002. 95-100

[12] Pankaj P, McKeown N. Packet classification on multiple fields[C]. Proceedings of ACM Sigcomm, Cambridge, 1999. 147-160

[13] Lakshman T V, Stiliadis D. High-speed policy-based packet forwarding using efficient multi-dimensional range matching[C]. Proceedings of ACM Sigcomm, Vancouver, Canada, 1998. 191-202

[14] Baboescu F, Varghese G. Scalable packet classification[C]. Proceedings of ACM Sigcomm, Califonia, 2001. 199-210

[15] Srinivasan V, Varghese G, Suri S, *et al*. Fast and scalable layer four switching[C]. Proceedings of ACM Sigcomm, Vancouver, 1998. 203-14

[16] Srinivasan V, Suri S, Varghese G. Packet classification using tuple space search[C]. Proceedings of ACM Sigcomm, Cambridge, 1999. 135-146

[17] Woo T Y C. A modular approach to packet classification: algorithms and results[C]. Proceedings of IEEE Infocom, Tel Aviv, Israel, 2000. 1 213-1 222

[18] Yao Xingmiao, Hu Guangming, Li Lemin. A multi-dimensional packet classification algorithm[C]. 2004 International Conference on Communications, Circuits and Systems, Chengdu, China, 2004. 670-673

编　辑　孙晓丹