自相似活跃子网前缀空间的路由查找

彭艳兵 龚 俭 丁 伟 徐加羚
(东南大学计算机科学与技术系华东北网络中心 南京 210096)

摘 要 IP 地址查询是路由器的基本工作,活跃 IP 和子网前缀地址空间是重尾分布且自相似的,而针对这种重尾分布的 IP 地址和前缀可以用于对路由查找进行统计优化. 文章分析并验证了活跃 IP 地址空间的特点和子网前缀空间分形自相似特性,活跃 IP 的子网前缀在不同的聚类规模上的次序统计量服从 Pareto 分布,主干路由表项的次序统计量也近似服从 Pareto 分布. 该文提出了一种基于活跃度排序的路由逐次查找算法——SOSL,对 IP 地址查询进行了优化,在该文的模拟实验中,活跃路由表的规模、刷新周期和活跃度判定下限间存在一些对数线性关系,使得作者可以以很小的活跃路由表来实现全部路由查找需求的 99 %;为 SOSL 实现中最关键的活跃路由表排序问题提出了一个基于计数器溢出的方案,复杂度为 $O(1)$. 对比发现该文的算法与 TCAM 结合能够提高 TCAM 的效率,高效地控制活跃路由表的规模,易于硬件实现.

关键词 活跃 IP;子网前缀;重尾分布;路由查询;统计优化;溢出排序
中图法分类号 TP393

Route Lookup in Fractal Active Subnet Prefix Space

PENG Yan-Bing GONG Jian DING Wei XU Jia-Ling
(*CERNET East China (North) Regional Center, Department of Computer Science and Technology, Southeast University, Nanjing* 210096)

Abstract Route lookup is a basic work for the routers. The active subnet prefix is self-similar at different prefix length, and this fractal distribution can be used to statistical optimizing the route lookup. After validating self-similar characteristic of the active subnet prefix space, it was found that the order statistic of the active subnet prefix at different aggregation class obeys to Pareto Distribution, and the active subnet distribution of different subnet prefix scale is self-similar. Further investigation disclosed that the active degree of route items in backbone router is close to Pareto distribution. Based on these results, this paper proposes a new route lookup method —— Statistical Optimized Successive Lookup algorithms (SOSL), which optimizes the route lookup by the active route sorting. The simulating experiment suggests that there are some logarithmic linear relationships among the scale of active route table, refreshing period of active route table and counter's overflowing value, and which enable router to classify the packet in a very small active route table. A fast sort scheme is presented to solve the key bottleneck, route active degree sorting, in SOSL with $O(1)$ time complexity metrics. The comparison among SOSL, Longest Prefix Matching algorithm (LPM) and TCAM implied that SOSL is easy to implement by hardware. It is the greatest advantage that SOSL can work along with TCAM to improve the efficiency of TCAM with efficient active route table scale management.

Keywords active IP; subnet prefix; pareto distribution; route lookup; statistical optimizing; overflowing sort scheme

收稿日期:2005-03-04;修改稿收到日期:2005-05-18. 本课题得到国家"九七三"重点基础研究发展规划项目基金(2003CB314803)、国家自然科学基金(90104031)资助. 彭艳兵,男,1974 年生,博士研究生,研究方向为网络行为学. E-mail:ybpeng@njnet.edu.cn. 龚 俭,男,1957 年生,博士,教授,博士生导师,主要研究方向为网络管理、网络安全、网络行为学等. 丁 伟,女,1962 年生,博士,教授,主要研究方向为网络管理、网络安全、网络行为学等. 徐加羚,男,1980 年生,硕士研究生,主要研究方向为网络测量.

1 引 言

路由器的基本工作是 IP 地址查找. 对于高速主干路由器,路由表的路由项一般在 150 K 左右,最高可达到 500 K 左右. 2. 5 Gbps 的主干网络已经普及,10 Gbps 的主干网正在广泛部署,40 Gbps 以上的线路已开始投入在商用. 由于路由器需要以 100Mpps 的查表速率才能满足 OC-768(40 Gbps)端口 48B 长度 IP 报文的线速转发,每报文的处理时间要求小于 10ns[1]. 高速的主干网络流量对路由表的查询提出了很高的要求,因此各种快速地址算法应运而生,用于减少缓存占用和加速查找操作. Henry 等人[2]指出,高速路由查找算法需要考虑的重点问题是更高的性能,或者更小的内存占用.

现有路由查找算法有很多,如最长前缀匹配树、线性比较查找、基于哈希的查找以及结合哈希和树的查找. 基于哈希的查找性价比太低,线性比较查询在时间和空间上消耗太多,无法利用快速 SRAM 等硬件带来的好处[2],TCAM 硬件算法的速度最高[2],但是其高昂的成本、低下的效率使得它只能应用于对速度非常挑剔的环境里.

Henry 等人[2]提出了一种基于最长前缀匹配的快速路由查找算法,查找快速且空间占用低;Sarang 等人[3]给出了最长前缀匹配基于 Bloom Filter 的改进,但是性能上的改进不大. 现在的高速路由算法已经把查找和比较次数降低到 3 次以内,如线性表查找法[4],TCAM[5,6]等.

使用缓存优化的路由查找有很多方案,很多都是基于 Hash 的缓存来加速查找速度,Chang 等人[6]提出一个基于 Bloom Filter 老化淘汰的缓存方案,并研究了它对性能的影响,在允许少量的误分类的情况下,该算法极大地减少了缓存的需要量;Shi 等[7]提出一个基于 Hash 负载均衡的并行报文分类系统的缓存方案,指出 Hash 的均衡性对缓存性能的影响极大,长流会使得并行系统的载荷颠簸;Chang 等人[8]提出基于摘要缓存的路由查找方案,在牺牲一定精度的前提下减少了缓存的大小;Li 等人[9]研究了路由查找缓存的三个重要方面,缓存的关联性、淘汰测量、Hash 函数的选择都对缓存的路由查找有不同程度的影响. Woo 提出[10]使用跳转表、查找树和 Filter bucket 组成的三层结构的路由查找算法,其中的跳转表使用了基于统计的权值调整.

这些路由查找算法中逐项比较查找,因其查找次数多、速度慢、空间占用率高而一直被人们所忽视,但是其算法的简洁却是其它算法所无法比拟的. 本文提出针对逐项比较查找的改进,不使用 Hash 函数,而以很少的代价实现和最长前缀匹配树算法相媲美的性能,与 TCAM 算法结合后还能够有更高的性能改进. 本文也主要针对与这些快速路由查找算法的比较来展开问题的讨论.

Kohler 等人[11]提出网络上可见的 IP 地址空间是多重分形自相似的,不同的站点甚至不同的聚类都有不同的分形特征,这些地址的结构能够用双参数的多重分形的康拓尘埃来很好地建模. 程光等人[12]通过对 CERNET①江苏省网边界主干进行观察,给出了活跃 IP 地址的非均匀分布,其中 80 %的流量集中在 15 个 IP 地址,99. 9 %的流量集中在 5000 个活跃 IP. 因此,我们可以利用活跃 IP 的非均匀分布来进行统计优化,经过优化的 IP 地址的查找预期可以与最长前缀匹配媲美. 因此本文提出一种基于统计优化的逐次比较路由的查找算法 —— SOSL (Statistical Optimized Successive Lookup),并将其与其它算法进行了比较.

本文用到的几个定义先在这里说明:

子网前缀空间:由子网前缀比特串构成的正整数空间,用于研究活动子网分布.

归一化的子网前缀空间:为了便于不同长度的子网前缀间进行比较,这里对子网进行了归一化处理,即子网前缀的值除以该子网前缀地址空间的大小,用于研究不同规模的活动子网情况和活跃子网的自相似情况,例如对于 B 类地址前缀,则是用 2^{16} 去除对应 IP 地址中的网络地址部分,因此不同前缀长度的两个活跃子网也可能会出现在归一化空间相同的位置上.

平均路由查找次数:一次路由查找由多次内存访问构成,因此路由表访问次数的平均值是路由项命中率的数学期望.

本文分 4 个部分阐述分形空间地址查找算法,第 2 节描述了活跃 IP 地址空间的分形特性;第 3 节提出一种基于活跃路由表的路由查找算法 —— SOSL,并就其关键问题进行了分析;第 4 节把 SOSL 算法与其它路由查找算法进行了对比,归纳了 SOSL 的优越性;第 5 节是本文的主要结论和将来的工作设想.

2 活跃 IP 地址空间分形特征

本文通过实验对各子网前缀和路由进行了统计,以其在单位时间总体中所占的比例,即命中率,

① CERNET:中国教育和科研计算机网(China Education and Research NETwork),http://www.edu.cn/

作为这些对象活跃度的标准. 通过对 IP 地址空间进行观察,从实验可以发现活跃自相似结构

本文的实验数据采集于 CERNET 江苏省网边界,采集于 2004 年 4 月的某天,主干带宽为 1000Mbps,日平均流量为 587Mbps. 4 小时的活跃 IP 被归纳到不同的子网规模,挑选出其中最大的 5000 个(Top5000)进行分析. 由于不同的子网前缀的规模不一样,其中短前缀是长前缀的合并和聚类,合并后的子网数量会减少,因而会有一些命中率非常小的活跃子网进入 Top 5000. 为了便于比较,这里只取命中率大于 1.0E - 5 的活跃子网进行研究,如下图所示.

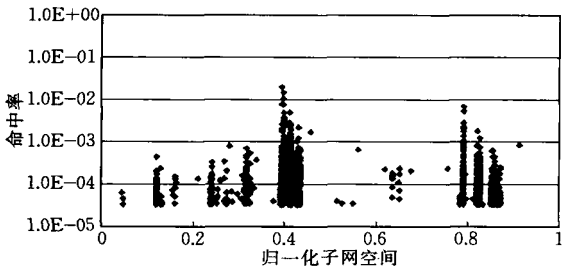

图 1　24 比特子网掩码时的活跃子网分布

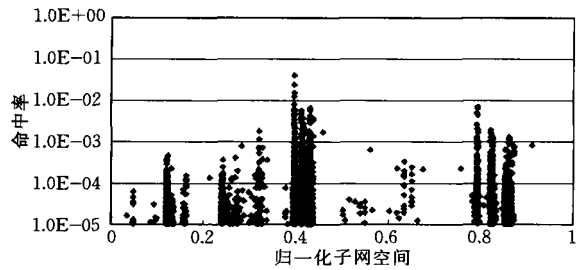

图 2　20 比特子网掩码时的活跃子网分布

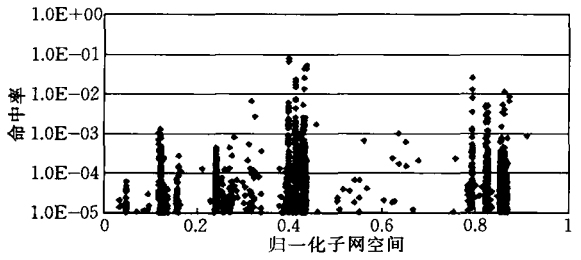

图 3　16 比特子网掩码时的活跃子网分布

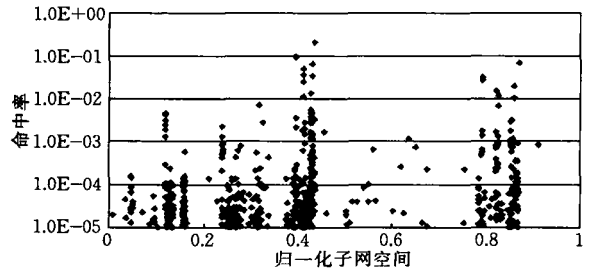

图 4　12 比特子网掩码时的活跃子网分布

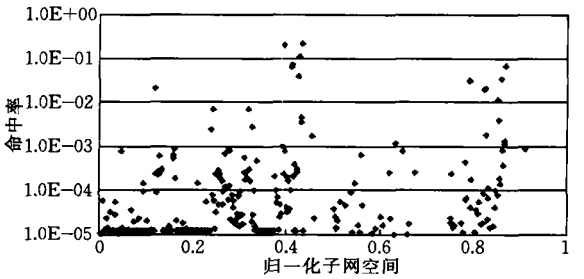

图 5　8 比特子网掩码时的活跃子网分布

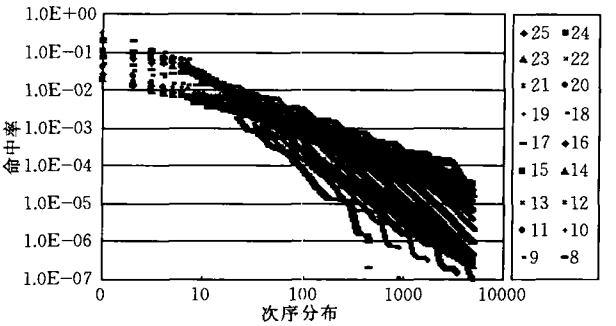

图 6　Top5000 子网活跃度的次序分布

图 1～图 5 列出了不同子网前缀的活跃子网分布,水平轴为归一化子网前缀空间. 从图中可以看出,占主要成分的活跃 IP 在子网前缀空间是稀疏的. 网络上可见 IP 的地址空间符合多重分形分布的康拓尘埃, Kohler 等[11]人对此进行了详细的描述. 但根据我们的观察,活跃的 IP 子网间的不平衡关系不能用简单的基于等概率模型的多重分形的康拓尘埃来描述,只能用经过裁剪的康拓尘埃来描述,而裁剪的规则可以看成是活跃 IP 和子网的特有分形特性.

图 1 的活跃子网分布可以看出,所有 Top5000 的活跃度都在 1.0E - 5 的基数上. 而当子网规模缩小时,如图 2～图 5 所示,有很多活跃度很低的子网进入了 Top5000. 而当子网规模缩小时, Top5000 里的低活跃度的子网越多,高活跃度的子网越来越集中. 如图 5 所示,活跃度高的子网数变得非常稀疏,而图 3～图 5 的变化过程也突出了这个规律,少数的活跃子网集中了大部分的流量. 从图 1～图 5 可以看出,不同前缀长度的子网间的活跃度分布是相似的,与子网的规模无关. 在前缀长度缩小的过程中,可以明显看到邻近子网的凝聚过程.

从图 6 的数据分析可以得出,子网活跃度的秩序统计近似服从 Pareto 分布; Pareto 分布的参数是不一样的,子网前缀越短,则子网的聚类越多, Pareto 分布的曲线下降越快.

我们在同样的实验条件下对 CERNET 江苏省主干路由器的动态路由表进行了观察.

图 7 是 CERNET 主干路由表子网前缀长度的次序统计,近似服从 Pareto 分布. 根据 Falconer 提出的观点,两种分形的交集还是分形[13],可以预计,这种不同长度的子网前缀的分形组合也服从 Pareto 分布,路由器里的路由活跃度分布就是这样的例子. 在这样的非均匀分布下,从 CERNET 主干路由器上采集到的 14995 项的路由表的路由命中率统计如图 8 所示.

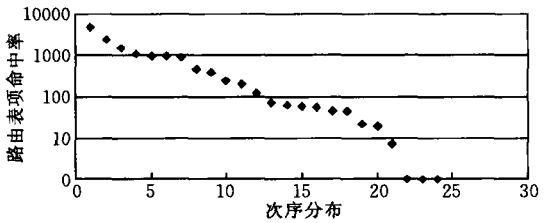

图 7 CERNET 主干路由表子网前缀长度次序分布

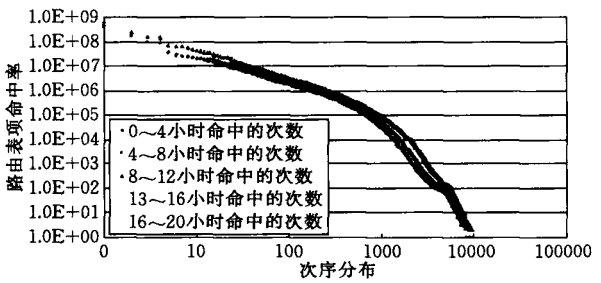

图 8 CERNET 主干路由命中数的次序统计

从图 9 中可以看出,路由表的表项的命中率并不严格服从 Pareto 分布,其前 600 项对 Pareto 分布服从的比较好,但是 600 项后下降的梯度比前面的大. 本次实验共观察了 20h,每个粒度 4h,这 5 条曲线的形状和位置比较接近,表明在不同时间段观察的曲线不会有太大的差别. 也就是说,这种重尾分布的路由表命中率在时间上是稳定的. 另外,对于 15K 的路由表项,在缓存容量大于 1K 项后,对性能的提升就很小了. 这样,我们以不到路由表总路由项数十分之一的 Cache 容量就获得了 99%以上的报文命中率.

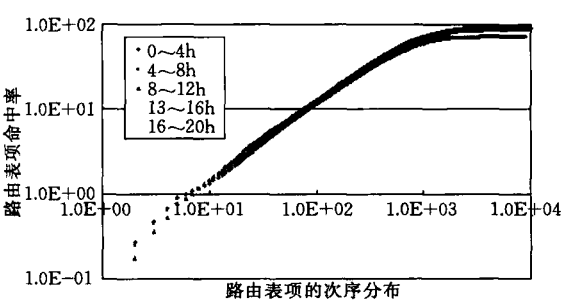

图 9 路由表平均查找次数 I 的分布

图 10 是 IPLS 公布的 Trace CLEV 的①活跃 IP 分布的次序统计和累积分布,可以看出对于总数为 506521704 个的活跃 IP 而言,只要处理了 5378 个高度活跃的 IP,就处理了 90%的报文;只要处理了 74488 个高度活跃的 IP,就处理了 99%的报文,显然如果把这些活跃 IP 归纳成路由,其活跃的路由表的规模将比这些活跃 IP 的数量小很多. 分形分布与 Pareto 分布间存在着一定的联系[14],在图 1～图 8 里得到了证明. 图 10 表明,其它网络里也存在着重尾分布和自相似行为. 下面更深入的研究表明,活跃路由表的规模和处理的报文数的规模以及活跃度的阈值定义有关.

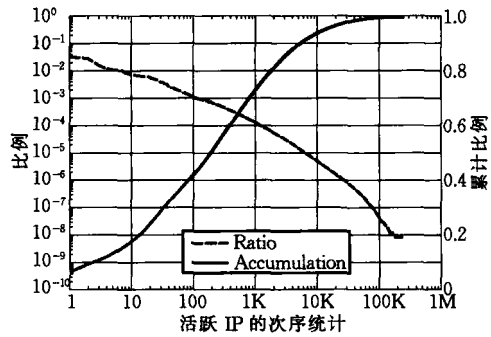

图 10 IPLS_CLEV 的活跃 IP 的重尾分布和其累积函数

由于活跃 IP 的分形分布导致的子网活跃度的重尾分布,可以对路由按照活跃度或者命中率排序,对排序后的路由表按照线性逐次查找的算法来查找路由,可以预计会有好的性能;对活跃的路由表进行并行查找也能够加速查找的速度. 下面将对这种逐次路由查找算法进行分析.

3 SOSL 算法的实现和关键问题——活跃度排序和与其它算法的结合

3.1 活跃度排序的相关问题

SOSL 算法实现是基于活跃路由表,如表 1 所示,结构图如图 11 所示. 在图 11 里,对于到达的 IP 查询,SOSL 算法对线性路由表进行逐项查找,找到后对该路由进行计数;另外有一个模块对计数器的值进行排序和清零. 排序的目的是优化线性路由表,

① IPLS CLEV 发布的 Trace,时间是 2002-08-14 09:00:00～11:00:00,OC48c(2.5Gbps) Packet-over-SONET,地址经过了净化处理,因而只能研究活跃 IP 的次序分布,http://pma.nlanr.net/Traces/long/ipls1.html

清零的目的是防止计数器溢出时打乱优化好的路由表. SOSL 算法可以分为如下的两个并行部分:

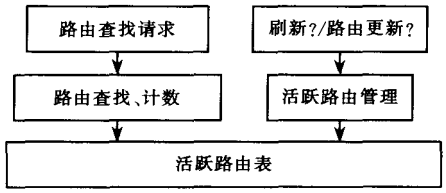

图 11　基于 SOSL 的路由查找框架结构

1. 路由查找

lookup in the Active Route Table
until (longest prefix *i* matched)
　　counter[*i*] + + ; *AllPacketCounter* + +
　　return Forward Direction;

2. 活跃路由表优化和计数器刷新

on event (AllPacketCounter >
　　Threshold or route update)
　　refresh the active route table;
　　Reset all counters to 0;

表 1　SOSL 算法的活跃路由表

i	Route Item	Forward Direction	Counter	LPM bit
1	Route/ mask 1	Next Hop 1	count1	off
2	Route/ mask 2	Next Hop 2	count2	off
3	Route/ mask 3	Next Hop 3	count3	on
…	……	……	……	

从上面可以看出,尽管算法的实现简单,但是也对其中的排序部分提出了很高的要求,因为查找已经是常数量级的计算复杂度,为了不增加整个算法的计算复杂度,排序算法也必须是常数量级的计算复杂度. 下面我们讨论如何实现计算复杂度为常数量级的路由表排序.

排序优化有快捷的算法,由于 SOSL 的每个路由都带有一个计数器,这些计数器会在某些时刻发生计数器溢出. 如果我们把这些计数器的值域看作是赛跑比赛的路程,则跑得最快的比赛者会先触线. 与赛跑触线者胜出的规则类似,本文利用计数器的溢出顺序作为路由优先顺序的判别,由于累加和交换排序的计算量小于 3,都是时间复杂度为 $O(1)$ 的算法,这样实现的计数器优先时间复杂度和空间复杂度都能够满足 SOSL 的要求. 本文把这种排序方法称为溢出排序法.

为了实现路由的最长匹配,在路由器里设置一个 LPM bit 的标志位,如表 1 所示. on 表示还有最长的匹配前缀,off 表示没有最长前缀匹配的要求,或者已经是最长的前缀匹配了. 同时在路由更新的时候: (1) 如果未找到匹配,插入新路由项; (2) 如果找到前缀匹配的 LPM bit 标志为 off,且掩码长度不大于当前掩码长度,直接更新;否则,设置 LPM bit 为 on,插入新路由项; (3) 如果匹配项为 on,需要继续查找; (4) 返回最长前缀匹配项.

在对普通报文进行路由查找的时候,需要判断 LPM bit 是否为 off,如果为 on,则需要继续查找,最后返回最长的路由前缀匹配. 若利用 TCAM 的并行查找功能,则这些逐项查找可以一次完成,同时也可以利用 TCAM 的 LPM 功能.

计数器的位宽和刷新时机的选择非常重要. 对于 40 Gbps 主干,要求报文处理能力在 100Mpps 以上,因此,选择每处理多少报文刷新一次计数器(刷新周期)可以通过需求来动态确定. 下面以从 Ripe 收集的 158231 的路由表[①]和从 CERNET 主干收集的 14995 K 的路由表,在图 1～图 6 里使用的 Trace 里进行路由查找的模拟,以说明计数器刷新的时机选择与 SOSL 活跃路由表规模、活跃判定标准的关系.

图 12 是 CERNET 14995 项路由表的规模、刷新时机与计数器位宽的关系,图 13 是 Ripe Amsterdam 158231 项路由表的规模、刷新时机与计数器位宽的关系,图例中的数字代表计数器的有效范围,则图 12 和图 13 表示的是命中率大于该数字时的活跃路由的数量,其以路由的命中率作为路由活跃与否的度量,阈值的下限为计数器的最大表示范围. 可见,如果每处理 100M 个报文刷新一次计数器,CERNET 路由表中的命中率大于 0 的活跃路由表的规模在 5. 6K 条,RIPE Amsterdam 路由表中的命中率大于 0 的活跃路由表的规模在 24. 7K 条左右. 上述活跃路由表处理了 100 %的报文. 而当计数器位宽达到 8bits,即当路由的活跃命中率下限为 256 的时候,每 100M 个报文刷新一次计数器,CERNET 活跃路由表的规模在 1. 4K 条,RIPE Amsterdam 活跃路由表的规模在 1. 6K 左右条. 如果选择每 1M 个报文刷新一次计数器(对应于每秒刷新 100 次计数器),则对于 4 比特位宽的计数器,CERNET 的活跃路由表的规模在 891 条,RIPE Amsterdam 活跃路由表的规模在 742 条. 此时,CERNET 的 891 条的活跃路由表处理了 99. 2 %的报文,RIPE Amsterdam 的 742 条活跃路由表处理

① http://www.ris.ripe.net/cgi-bin/rrcstatus.cgi, Amsterdam, 收集 2005-02-18 00:00:00 的路由数据

了 99.0 %的报文.如果选择每 1M 个报文刷新一次计数器(对应于每秒刷新 100 次计数器),则对于 8 比特位宽的计数器,CERNET 的活跃路由表的规模在 270 条,RIPE Amsterdam 活跃路的规模在 134 条.此时,CERNET 的 270 条的活跃路由表处理了 94.0 %的报文,RIPE Amsterdam 的 134 条活跃路由表处理了 94.8 %的报文.显然,使用 4bit 的活跃计数器,每 1M 报文刷新一次计数器,可以保证 150 K 路由表中的活跃路由表能够处理 99 %以上的报文.活跃路由表的规模、刷新周期与计数器位宽间近似的对数线性关系可以从 Pareto 分布推导出来.

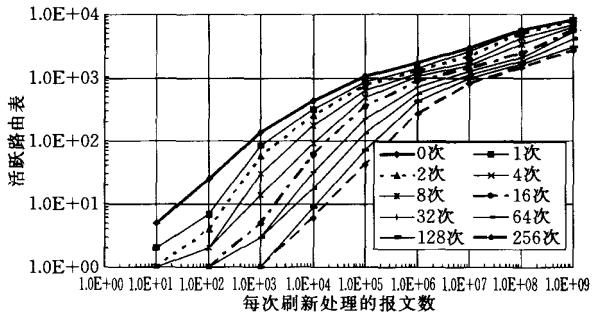

图 12 CERNET 活跃路由表规模、刷新时机与计数器位宽

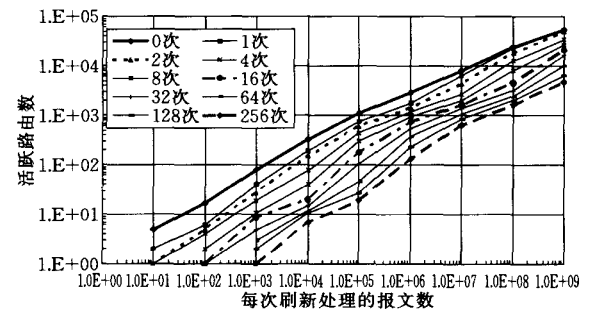

图 13 Ripe Amsterdam 活跃路由表规模、刷新时机与计数器位宽

在刷新活跃路由表的时候,如果活跃路由表里不活跃的项目太多,可以增加每次刷新前处理的报文数量;如果不活跃项目太少,或者全部都活跃,表明每周期处理的报文数太多,应该缩小相应的刷新周期,这样就可以根据报文的实际动态分布实现刷新周期的动态调整.这个动态调整周期功能特别有利于路由器开始工作的初期迅速构建活跃路由表.

更新活跃路由表的开销在整个报文查找过程中的比重可以计算出来,假设对于 100Mpps 的报文查找速度,每个刷新周期处理 1M 个报文,每秒共需要刷洗活跃路由表 100 次;4bits 的活跃计数器可以保持活跃路由表在 1 K 以下,每更新一条记录相当于一次报文查找过程,则更新的计算复杂度与 100M 报文相比,相当于处理了 100 ×1 K = 100 K 个报文,即刷新活跃路由表只有不到 0.1 %的负荷,显然是可以接受的.

3.2 与其它算法的结合

由于 SOSL 处理了 99 %的需要路由查找的报文,所以与 SOSL 结合、处理剩余 1 %查找工作的路由查找算法并不要求有很高的性能,唯一的要求就是其在 SOSL 进行活跃路由表刷新时能够提供 SOSL 所需要的活跃路由.根据 3.1 节里的描述,该算法只需要为每个有效的路由项配置一个 4bit 的计数器,并且在计数器溢出的时候能把溢出的路由项作为活跃路由进行管理和记录;在活跃路由表刷新的时候,替换活跃路由表里不活跃的路由;在活跃路由表进行刷新以后把其所有的计数器清零即可.下面给出一个结合了 TCAM 和基于 Trie 的分支查找的 SOSL 查找的例子.

图 14 是结合了 TCAM 和基于 Trie 分支查找[2]的完整的 SOSL 路由查找算法. TCAM 使用了一个 4bit 的计数器来记录不活跃的路由表项,这样 TCAM 被 SOSL 改造成了一个缓存,缓存的调度机制由活跃路由管理单元实现;TCAM 把 SOSL 的逐项查找变成了并行查找,极大地提高了 SOSL 的速度,同时 SOSL 也可以利用 TCAM 的 LPM 功能来实现最长前缀匹配;基于 Trie 的分支查找算法也增加了一个 4bit 的计数器,用于找到最活跃的路由分支,其包含了完整的路由表.活跃路由管理单元负责在活跃路由刷新的时候把 TCAM 里的不活跃路由表项与 Trie 查找算法里的活跃表项进行交换,并且在路由更新的时候负责刷新路由表;Reset 的目的是使所有的计数器清零,为下一个活跃路由刷新前的处理周期做准备.活跃路由表里的 LPM Label 可以用来压缩活跃路由表里的 LPM 树,将不在本文进行详细研究.

4 与其它路由查找算法的比较

求平均路由查找次数的公式根据概率理论,设到达的路由查询访问次数为 n,在路由表 N 内找到路由的概率是 1,i 为某路由的路由表查找次数,如下计算:

$$I = E(i) = \int_N \times p(i)\,\mathrm{d}i = \left(\sum_i^N N_i \times i\right) \Big/ N = \sum_i^N p(i) \times i \qquad (1)$$

也就是路由访问概率与路由访问次数的积的和.

4.1　没进行统计优化的路由表平均查找次数 I

假设对没有优化的路由表的访问服从均匀分布,即 $p(i) = 1/N$,则第 i 项被访问的概率为 $1/N$;对于 n 次 IP 访问:

$$I = \frac{1}{n}\sum_{i=1}^{N} n \times p(i) \times i = \sum_{i=1}^{N} \frac{i}{N} = \frac{(N-1)}{2} \qquad (2)$$

当有 n 次 IP 访问时,路由器需要查找 $n(N-1)/2$ 次路由表,计算量巨大.

4.2　LPM 和 TCAM 的路由表平均访问次数

对于最长前缀匹配,没有线性路由表可以供我们使用,但是根据最长前缀匹配的机理,可以推出,若假定这种算法是在整个路由表范围内路由的命中率是均匀分布的,对于每个到达的报文,最长前缀匹配的最大路由表访问次数是 32,最小路由表访问次数是 8,可知其平均访问次数为 20,则平均每个报文最长前缀匹配需要访问 20 次路由表. 但是根据图 7,各长度不同的前缀在路由表里出现的频率是不同的,如果使用前缀长度在路由表出现的频数作为前缀长度命中概率的衡量,可以用平均子网前缀长度作为平均路由查找次数,如图 7 中的平均子网前缀长度为 21.3.

TCAM 由于只需一次路由表访问,则路由表的平均访问次数为 1. 但是 TCAM 在解决最长前缀匹配问题的时候与 SOSL 一样烦琐[15].

4.3　统计优化的路由表

一般的路由器使用最长前缀树匹配[2],最多只需 32 次查找,其它路由查找算法都是常数量级的查找次数,如果我们的算法能够使得平均查找次数与最长前缀匹配的平均查找次数在同一个数量级,就可能有比其它路由查找算法更好的性能,或者与其它算法结合后可能有更好的性能. 对于一个统计优化的线性路由表,我们在查找比例为 d 处优化截止,若路由表项命中的次序分布为 $f(i)$,此时对应的查找表为 L;剩下 $(1-d)$ 的查找按照普通的路由查找算法,其统计分布是 $p(i)$:

$$I = \frac{1}{n}\left(\sum_{i=1}^{L} n \times f(i) \times i \times d + \sum_{i=L}^{N} n \times p(i) \times (1-d) \times i\right)$$

$$= \sum_{i=1}^{L} i \times f(i) \times d + \sum_{i=L}^{N} i \times p(i) \times (1-d) \qquad (3)$$

如果只使用 SOSL 对全部路由表项进行查找,则 $L = N$, $d = 1$,上述计算公式退化为公式(4):

$$I = \frac{1}{n}\left(\sum_{i=1}^{N} n \times f(i) \times i\right) = \sum_{i=1}^{N} i \times f(i) \qquad (4)$$

图 9 列出了本文实验使用 SOSL 算法的路由平均查找次数分布,使用 CERNET 主干的路由表,水平轴为 L,纵轴为 I,采用的部分/全部路由表的平均查找次数,路由表按照路由命中率由大到小排序,使得线性查找时先查找访问命中率高的路由. 从图中的曲线可以看出,即全部的路由查找都使用 SOSL 线性优化来进行查找,平均访问次数都小于 100. 另外从图中还可以看出,在线性优化查找的路由表路由数维持在 1000 左右时,路由平均查找次数刚好在一个拐点,拐点左边平均路由查找次数与 SOSL 的规模的对数成正比,拐点右边的平均路由查找次数则维持在接近水平的位置. 此时取 $L = 1000$,活跃路由表处理了 99.2 %以上的报文. 图 13 里为了完成 150 K 路由表的 99 %的报文查找,每处理 100M 报文刷新一次活跃路由表,活跃路由表的规模也只有 24.7 K,每处理 1M 报文刷新一次活跃路由表,活跃路由表的规模也只有 3.7 K. 如果提高活跃路由的判定标准,活跃路由表的规模还可以减小.

4.4　SOSL 与其它算法的比较

现在我们使用一个路由表规模为小于 1 K 的活跃路由表,适当的活跃路由表刷新周期,就能够完成规模为 14995 的 CERNET 主干路由表 99.2 %的工作量,平均查找次数也维持在一个较低的水平,使用 TCAM 硬件算法实现并行查找还能够把 SOSL 算法性能提升到 2 次查找的水平;实现 150 K 以上规模的路由查找,也只需要加快活跃路由表的刷新周期,提高活跃路由的下限即可保持活跃路由表在很小的规模,如 3.1 节所述. SOSL 是基于统计模型,各部分的代价是可加的,因此可以将剩下不到 1 %的工作量由其它节省内存或者计算量的算法完成,能够达到更好的效果. 由此我们可以构造出一个能够与其它高速路由查找算法媲美的路由查找算法,内存的耗费很小,并且能够与其它路由查找算法兼容. 下面我们就比较一下这几种路由查找算法的优缺点.

SOSL 的一个优点是可以利用溢出排序法维持较小的活跃路由表规模;另外一个优点同时也是缺点是它必须与其它路由查找算法结合,使得这个基于 SOSL 的算法显得复杂. 但是它可以在与 TCAM 结合后,让 TCAM 像缓存一样工作,提高 TCAM 的效率;使缓存像 TCAM 一样并行查找,充分利用硬件的能力.

表 2 统计优化线性查找与最长前缀匹配查找的比较

	统计优化顺序查找 SOSL	最长前缀匹配	前缀排序的 TCAM
算法	顺序查找,统计优化	根据最长前缀匹配树二分查找	并行比较
查找次数	小于某个常数 I	最多 32	1
内存消耗	99 %的工作量可以在 $L \times H$ 内存里完成	$3 \times N \times H$	小规模适用
计算量	小于某个常数, $I+3$	查找次数小于 32 次生成前缀匹配树: $N \times H$	1
路由表更新	与路由查找过程一致	有生成和删除前缀匹配树的操作	多次移动路由, $W/2$ (PLO-OPT)[15]
优点	算法结构简单、快速,极大地减少了活跃路由表的规模,并且可以根据实际情况进行调节,易于硬件实现	稳定性好	快速
缺点	需要额外的装置来保持活跃路由表最新;需要与其它算法结合,解决剩下的 1 %的报文查找的问题;需要解决最长前缀匹配的问题	算法复杂,计算复杂度大,内存占用高,不易硬件实现	昂贵,内存 SRAM 的规模限制了路由表规模;更新时路由表移动次数过多;内存和计算的效率低下;需要解决最长前缀匹配的问题

5 结论和将来的工作

本文阐述了活跃 IP 地址分形对子网活跃度的影响,验证了子网活跃度和路由活跃度近似服从 Pareto 分布. 根据这个结论,提出一种路由查找算法——统计优化的线性查找法(SOSL). 本文发现 SOSL 与最长前缀算法和 TCAM 算法相比的优越性,SOSL 算法的实现简单,内存占用小,计算复杂度属于常数量级,易于硬件实现,这些都是最长前缀匹配算法和 TCAM 算法无法比拟的;而 SOSL 能够与 TCAM 和其它算法结合,实现小规模活跃路由表里的 TCAM 和基于新型活跃度调度的缓存路由查找方案.

本文还探讨了 SOSL 实现中比较重要的一个步骤——活跃路由项排序算法,提出一种计算复杂度为常数量级的算法——溢出排序法,利用累加计数器的溢出来对路由进行活跃度排序,在找出活跃和不活跃的路由的时候有非常理想的效果. 对于计数器位宽的选择、计数器清零的时机选择和活跃路由表规模的关系也进行了探讨,发现活跃度的阈值下限(即计数器的溢出点)越大,活跃路由表的规模越小;活跃路由表和计数器刷新期间内所处理的报文数越多,活跃路由表的规模越大. 刷新周期也可以根据活跃路由表里不活跃的路由的数量进行动态调整.

将来的工作是把现有路由协议与 SOSL 的结合,能够把路由活跃信息传递到另外一个路由器,以帮助刚开始工作的路由器构建自己的活跃路由表;在活跃路由表里有 LPM 的时候,如果把其所有的 LPM 路由分支都保留在活跃路由表里会使活跃路由表的规模变得非常庞大,如何在保证 LPM 的前提下,利用图 14 中活跃路由表里的 LPM Label 减少活跃路由表的规模,是另外一个可以深入探讨的问题. 合并 LPM 各项中出口路由方向一致的路由以减少活跃路由表的规模也值得细致研究.

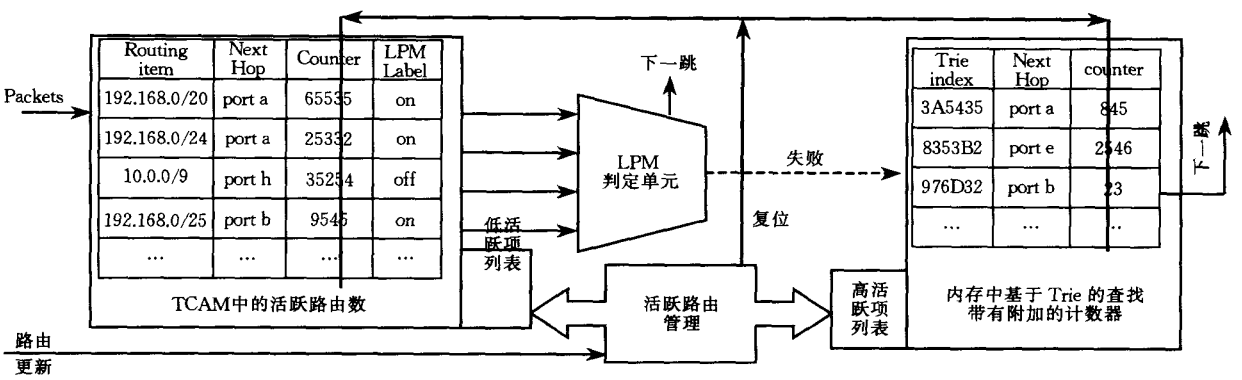

图 14 一个完整的 SOSL 算法的例子(结合了 TCAM 和 Trie 路由查找算法)

参 考 文 献

1 Wang Zheng-Xing, Zhang Yan-Xiao, Sun Ya-Min *et al*.. High-performance routing lookup based on prefix-range bi-search. Chinese Journal of Computers, 2004, 27(5): 604～610 (in Chinese)
(王振兴,张彦肖,孙亚民等. 基于前缀范围对分搜索的高性能路由查找. 计算机学报, 2004, 27(5): 604～610)

2 Henry Hong-Yi Tzeng, Tony Przygienda. On fast address-lookup algorithms. IEEE Journal on Selected Areas in Communications, 1999, 17(6): 1067～1082

3 Sarang Dharmapurikar, Praveen Krishnamurthy, David E. Taylor. Longest prefix matching using bloom filters. In: Proceedings of SIGCOMM Conference, Karlsruhe, Germany, 2003, 201～212

4 Wang Zheng-Xing. NGI-high-performance routers' forwarding process algorithms & implementations [Ph. D. dissertation]. Nanjing University of Science & Technology, Nanjing, 2004 (in Chinese)
(王振兴. NGI 高性能路由器转发处理算法与实现[博士学位论文]. 南京理工大学,南京, 2004)

5 Rina Panigrahy, Samar Sharma. Sorting and searching using ternary CAMs. IEEE Micro, 2003, 23(1): 44～53

6 Chang Francis, Feng Wu-Chang, Li Kang. Approximate caches for packet classification. In: Proceedings of ACM SIGCOMM (poster session), Karlsruhe, Germany, 2003, 4: 2196～2207

7 Shi W., MacGregor M. H., Gburzynski P.. Effects of a Hash-based scheduler on cache performance in a parallel forwarding system. In: Proceedings of CNDS'03, Orlando, Florida, 2003, 130～138

8 Chang Francis, Feng Wu-Chang, Feng Wu-Chi, Li Kang. Efficient packet classification with digest caches. In: Proceedings of HPCA NP3 Workshop, Madrid, Spain, 2004, 13～24

9 Li Kang, Chang Francis, Burger Damien, Feng Wu-Chang. Architecture for packet classification caching. In: Proceedings of IEEE ICON, Sydney, Australia, 2003, 111～117

10 Woo T. Y. C.. A modular approach to packet classification: Algorithms and results. In: Proceedings of IEEE Infocom2000, San Francisco, CA, 2000, 1210～1217

11 Kohler Eddie, Li Jinyang, Paxson Vern, Shenker Scott. Observed structure of addresses in IP traffic. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement, Marseille, France, 2002, 253～266

12 Cheng Guang, Gong Jian, Ding Wei. Research on the distribution of activated IP flow in the large-scale networks. Chinese Computer Science, 2003, 30(4): 43～46 (in Chinese)
(程　光,龚　俭,丁　伟. 大规模互联网活动 IP 流分布研究. 计算机科学, 2003, 30(4): 43～46)

13 Falconer K. J.. Fractal Geometry ——Mathematical Foundations and Applications. Wiley, 1990
(肯尼斯·法尔科内. 分形几何——数学基础及其应用. 曾文曲等译. 东北大学出版社, 2001, 135)

14 Kihong Park, Walter Willinger. Self-similar Network Traffic and Performance Evaluation. Wiley-Interscience Publication, by John Wiley & Sons, Inc., 2000, 531～553

15 Shah Devavrat, Gupta Pankaj. Fast updating algorithms for TCAMs. IEEE Micro, 2001, 21(1): 36～47

PENG Yan-Bing, born in 1974, Ph. D. candidate. His research interests focus on network behavior.

GONG Jian, born in 1957, Ph. D., professor, Ph. D. supervisor. His research interests include network management, network security, network behavior etc.

DING Wei, born in 1962, professor, Ph. D. supervisor. Her research topics involved network management, network measurement, network behavior etc.

XU Jia-Ling, born in 1980, M. S. candidate. His research interests focus on network measurement.

Background

The project 2003CB314803 studies the dynamic behavior in network. It is a subproject of the National Basic Research Program of China (also called 973 Program) 2003CB314800 which focuses on the theory for the new generation architecture of Internet. By network measuring, the authors are involved in disclosing the theory basement of network behavior and expanding its application. The measuring and metrics' theory and application is the major direction of our group, e. g. Network Traffic Sampling Measurement Model on Packet Identification by Cheng Guang *et al*. published by Chinese Journal of Computers in Vol. 26 No. 10. This paper is a application of the network packet dynamic distribution, which can optimize the high-speeded packet classification for core router designing.