接入网MAC层QoS系统多维报文分类算法的研究与应用

张晓彤[1] 李培娅[1] 宋丽华[2]

(北京科技大学信息工程学院 北京 100083)[1] (北方工业大学信息工程学院 北京 100044)[2]

摘 要 以HFC网络核心设备双向CM(Cable Modem)为研究背景,首先对报文分类经典算法和最新算法研究进展进行总结和分析,然后依据HFC网络QoS系统需求提出了一种基于B树结构和无冲突Hash函数的BH报文分类算法,并给出了该算法的详细设计和实现过程。通过理论分析得出该算法具有时间复杂度较低和占用内存小的特点,适合于CM等嵌入式应用环境。

关键词 报文分类算法,QoS,分类器,HFC

High-dimensional Packet Classification Algorithm Research and Application in MAC QoS System of Access Network

ZHANG Xiao-tong[1] LI Pei-ya[1] SONG Li-hua[2]

(Information Engineering School, University of Science and Technology Beijing, Beijing 100083, China)[1]

(Information Engineering School, North China University of Technology, Beijing 100044, China)[2]

Abstract Based on the study of HFC network core equipment CM (Cable Modem), this paper made a summary and analysis with the classical and the latest research progress of the packet classification algorithm; and then in order to meet the requirement of HFC network QoS system, a BH packet classification algorithm was proposed, which is based on B-tree structure and non-conflict Hash functions, and the design and implementation process was given out. Theoretical analysis shows that the proposed BH algorithm has less time complexity and small memory occupation, which is suited to embedded system applications such as CM and so on.

Keywords Packet classification, QoS, Classifier, HFC

1 引言

报文分类是指将报文与事先规定的规则进行比较,分类到相应流/服务的过程。事先规定的规则基于一个或多个报文头字段(或报文内容),称其为分类规则,所有规则的集合称为分类器。每个分类规则关联一个行为,以便对符合该规则的报文做相应的处理或标记。

报文分类在防火墙、入侵检测、路由、QoS(Quality of Service)等网络技术领域有着广泛的应用。在虚拟专用网(Virtual Private Network, VPN)中,服务提供商在路由器中根据报文头的源地址和目的地址字段将用户划分为不同的通信域,每个通信域构成一个VPN。包过滤型防火墙技术根据数据包头源地址、目的地址、端口号和协议类型等标志确定是否允许通过,从而有效保护网络的内部安全,入侵检测系统则利用报文分类和记录来跟踪传输对象。在服务质量系统中,为预先定义的规则进行流标记,用分类算法将每个报文分类到相应的流,根据流所定义的QoS参数集进行优先级排序和调度来保证该业务质量。总之,像防火墙、入侵检测、路由、QoS等网络技术都要求对基于报文头或者报文内容的某些字段进行报文分类,使得报文分类成为众多网络应用的关键技术之一,其分类速度的快慢、功能的强弱、存储空间的大小都直接影响到所在网络的整体性能。

本文首先对典型的报文分类算法和最新动态进行了研究,并给出初步的分析结果。通过分析这些算法的适用环境、性能优势以及存在的问题,以HFC网络QoS分类器为研究对象,设计了一种BH报文分类算法,并给出了该算法的详细设计过程和性能分析。本文结论对其它网络应用也具有借鉴意义。

2 报文分类算法最新研究进展

报文分类算法按实现手段可以分为两种:一种是硬件算法,另一种是软件算法,分别有不同的典型算法,如表1所列(其中n为规则的个数,d为分类的维数,w为分类字段的域宽)。

表1 典型算法[1]

分类	算法分类	经典算法	时间复杂度	空间复杂度	
软件	基本数据结构	线性查找算法	n	n	n
		Hierarchical tries 算法	wd	ndw	d2w
	几何算法	Grid-of-tree 算法	w^{d-1}	ndw	高
		AQT 算法	w	nw	—
		RFC 算法	d	n^d	高
	启发式算法	Hierarchical cuttings 算法	d	n^d	低
		Tuple-space search 算法	n	n	高

到稿日期:2008-04-01 本文受国家"八六三"高技术研究发展计划基金项目(2006AA09Z115),北京市科技产业化项目"SOC设计服务及重点产品关键技术研究"课题(编号D0306008041021)资助。

张晓彤 博士,副教授;李培娅 硕士研究生;宋丽华 博士。

	TCAM 算法	1	n	—
硬件	Bitmap-intersection 算法	n	dn^2	—

最新研究的算法大部分是基于以上经典算法的改进，另外也出现了一些新颖算法，本文以下部分将对这些算法进行详细介绍。

2.1 改进算法分析

HICNCH 算法[2] 基于 Hierarchical Intelligent Cuttings 算法结合源/目的端口号和协议类型的无冲突 Hash 函数，优点是搜索时间短、内存耗费低，具有 O(Base+Depth)时间复杂度(Base 定义为在到达 cutting tree 之前的内存访问次数，Depth 定位为 cutting tree 的深度)，缺点是规则更新复杂。

文献[3]中给出了一种 RFC 改进算法，结合多模式匹配算法的思想，对 RFC 算法进行了有益扩充，使新算法能够根据变长字符串域进行分类。源 IP、目的 IP、源端口和目的端口域按照 RFC 算法处理，为它们分别建立索引子表。对于两个字符串类型的域(如源/目的 IP 地址)，则结合多模式匹配算法的思想，时间和空间复杂度均为 $O(m)$(m 是字符串域包含的字符总个数)。该算法的局限性在于当字符串中包含的字符数较大时，需要较大的存储空间。

Bit Compression 算法[4]基于多维度范围查找的方法，解决 Bitmap-intersection 算法内存爆炸(memory explosion)的问题，具有 $O(dn^2)$～$O(dn\log n)$空间复杂度，需要额外的预处理时间。

Adaptive Rules Cutting 算法[5]类似于 HiCuts 和 HyperCuts 算法，将多维空间裁剪为更小的部分来实现规则的压缩，克服了 HiCuts 和 HyperCuts 算法的缺点，具备考虑所有维度的灵活性，但是需要在规则的更新速度上做进一步的研究。

文献[6]在 PCBNP 算法、RFC 算法和区域分割算法的基础上提出的一种新的可扩展的多维报文分类算法，具有 $O(d)$ 和 $O(d\times n)$的时间空间复杂度。第一阶段的压缩操作类似于 RFC 算法的第一次递归，但不采用启发式算法，而是以分析分类字段的特征为主，分析规则分布的统计规律为辅，采用固定的压缩算法；第二阶段不采用类似 RFC 算法的递归方法，通过采用索引列表结构表这一数据结构，减少存储空间，支持大的规则集，并用于网络智能小区建设。

CPHTIT 算法[7]结合 CrossProduct 和带索引表的 Hash 树算法，来满足快速 IP 分类的需求，其时间复杂度小于 $O(w^{d-1})$，适用于 IP 分类。

2.2 新颖算法分析

RC-FST 算法[8]适用于硬件实现，时间和空间复杂度分别为 $O(\log n)$和 $O(n^d)$。通过 IP 前 8 比特前缀建立 hash-compression 索引表将规则集拆分为多个子集，并为每个子集建立搜索树，易构建、优化和更新，搜索速度提高，缺点是内存消耗大。

预分离策略表的并行分类算法[9]，采用预分离模式消除规则的重叠，最初规则集预分离为多个子集，将子集的交集部分删除以减少规则的重叠，并结合 Quarter-Cut 决策树的并行分类算法，将大量规则存储在多个搜索引擎中，并行执行分类操作，具有 $O(\log n)$的时间复杂度。

文献[10]中提出了一种小内存消耗算法，运用独立集合的概念，内存消耗很少，搜索速度与规则集的个数和字段通配符的个数百分比无关，并行处理，算法在一般情况下更新处理很快。还有一种基于 Hash 的分类算法[11]，该算法利用每一维比特位分布的特点，通过遍历优化的查找路径以获取最佳匹配规则，包括 3 个查找过程：启发式的 Hash 查找、多路径动态搜索前缀查找和小规则集的查找。

2.3 性能测试方法

不论哪种分类算法，其实际性能测试都需要大量的规则集、数据报文以及这两者之间的匹配关系。要模拟产生这些有关联的数据集，本身是一项艰难的工作。目前华盛顿大学提出的基本测试工具 ClassBench[12]得到了广泛应用，这个基准测试程序基于 Internet 服务商、网络设备生产商等提供的 12 个真实规则集，规则的个数范围从 68 变化到 4557，主要分为 3 类：ACL，代表防火墙、企业的边缘/骨干路由器的接入链分类规则文件和包头文件；FW，代表防火墙的安全过滤规则集；IPC，代表基于软件系统的 VPN 和 NAT(Network Address Translation)过滤集。ClassBench 提供了规则集产生器、包产生器等工具，通过对参数的配置可产生自己所需规则集的规模。支持增量更新的 Markers-based Space Decomposition 分类算法[13] 以及 $O(\log W)$多维分类算法[14] 等均将 ClassBench 作为性能测试工具。但是考虑到报文分类多层次的概念以及分类维数的扩展性，ClassBench 以后的研究方向应该是提供 MAC 层以及对 IPv6 规则数据集的支持。

3 CableModem(CM)系统报文分类算法的设计

3.1 报文分类在 QoS 系统中的应用

本文以 HFC 网络核心设备双向 CM(Cable Modem)为研究背景，基于 DOCSIS(Data-Over-Cable Service Interface Specifications)规范[15]，在龙芯 CPU、SDRAM 内存芯片、嵌入式 Linux 操作系统等组成的嵌入式双向通信平台上实现 CM 与 CMTS(Cable Modem Termination System)之间 IP(Internet Protocol)的双向透明传输。在 HFC(Hybrid Fiber-Coax)网络中，CM 和 CMTS 将穿越射频接口 RF (Radio Frequency)的 MAC 帧分类为不同的业务流，并根据预先为该业务流定义的 QoS 参数集进行流量整形、策略处理和优先级调度等来保证业务质量。业务流、分类器是 Docsis 规范中业务质量的两个重要概念。业务流是一种 MAC 层的传输业务，提供 CM 发送的上行数据包或 CMTS 发送的下行数据包的单向传输业务，由一组诸如延迟、抖动及吞吐量保证等的 QoS 参数来表征。分类器是一组匹配规则，它将进入电缆网络的每个数据包与分类器规则进行匹配，以便确定该数据包被送到哪个业务流上进行传输。图 1 给出了 DOCSIS MAC 层内部的分类。

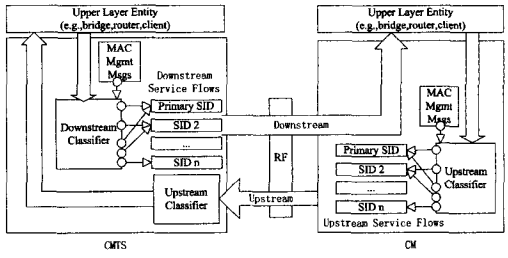

图 1 MAC 层分类器

CM 和 CMTS 的数据包分类由多个分类规则组成。多个分类规则有可能指向同一个业务流，分类器和业务流的关系如图 2 所示。如果发现分类器中某个规则的所有参数完全与

数据包匹配，则该分类器必须将数据包送到对应的业务流上，否则该数据包被分类到基本业务流(默认业务流)。

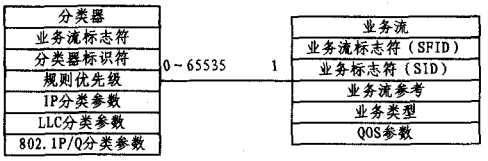

图2 分类器与业务流的对应关系

DOCSIS1.1规范定义分类器规则包括以下字段，如表2所列。在DOCSIS3.0版本中规则的分类字段还包括IPv6通信流类别、流标识、下一包头、源以及目的地址等。

表2 分类器规则字段

字段	说明
优先级	确定规则的搜索顺序
IP分类参数	零或者多个的IP分类参数(IP TOS Range/Mask，IP协议，IP源地址/掩码，IP目的地址/掩码，TCP/UDP源端口起始点，TCP/UDP源端口终点，TCP/UDP目的端口起始点，TCP/UDP目的端口终点)
LLC分类参数	零或者多个的LLC分类参数(目的MAC地址，源MAC地址，以太网类型/DSAP)
IEEE802.1P/Q参数	零或者多个的IEEE参数(802.1P优先级范围，802.1Q虚拟局域网识别符)
业务流标志符	代表该数据包被传送到那个特定的业务流

分类器既可以通过管理操作(配置文件、注册)也可以通过动态操作(动态信令DSX，DOCSIS MAC子层业务接口)实施规则的创建、修改或者删除。

3.2 BH算法设计与分析

报文分类算法的设计要遵循速度、内存和规则更新之间的性能折中原则，满足实时需要原则，遵循算法的简单性原则。在设计分类算法时，要充分考虑系统中分类器的特性，针对其特性采用规则个数的压缩、分类域宽的压缩、增加预处理时间等来加快报文分类的速度。

在本系统中，报文分类算法应用于分类器的分类过程，依照规范分析CM上行分类器特性可以发现：

1)在同一分类器中，许多不同规则的一系列字段存在重叠，造成空间的浪费。

2)由于多个分类器可以对应一个业务流，因此导致冗余规则的出现，减慢了分类的速度。

3)在进行字段匹配操作时，有基于IP地址的前缀查找、基于端口号的范围查找以及精确匹配，其中前缀表示、范围表示和精确值可以相互转换。

4)在嵌入式环境下，分类器应占用尽可能少的空间，并且支持增量更新。

5)在实际环境中，规则主要集中在IP层分类字段，LLC和802.1p/q分类字段相对较少。

由于目前大量的分类算法都集中在对IP层5维分类字段的分析之上，而CM系统的分类器维数达到了12维，并且现有算法对规则的增量更新支持较少，因此结合CM分类要求和嵌入式系统资源有限的特点，本文提出了采用B树结构结合Hash函数的报文分类算法，我们称其为BH报文分类算法。BH算法的基本思想是通过将规则集的复用字段存储在同一空间来减少内存的消耗，同时利用Hash函数来加快报文的分类，缩减规则集，并根据优先级策略获得最终业务流，分类过程如图3所示。

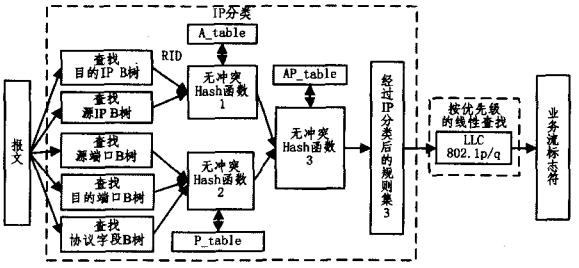

图3 报文分类处理过程

3.2.1 IP地址的分段结构

首先将IP地址的前缀表示转化为范围表示，得到IP地址的起点和终点。按照从小到大的顺序排列，这样就把整个地址分为m段。按照顺序对每个段分配一个连续的范围ID值(RID)，如图4所示。设范围起点为Sx，范围终点为Ex。每个RID与那些落入该段的clsID(分类器ID或者某个规则记录的地址)集合相对应，各个端点按B树结构存储。算法通过B树查找，为一个地址x找到小于等于x的最大前缀端点，从而得到其对应的范围RID。

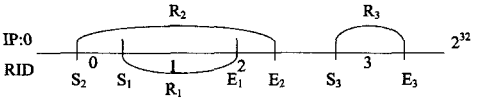

图4 IP段划分

3.2.2 无冲突Hash查找

源/目的端口采用的是范围表示，分段方法与IP字段类似。协议字段是精确匹配的，直接采用B树结构，记录协议对应的分类器ID即可。利用无冲突的Hash函数[16]进行规则集的查找：

$$H(s,d,p)=s\times D_{port}\times P_{proto}+d\times P_{proto}+p \quad (1)$$

其中$0\leqslant s\leqslant S_{port}-1, 0\leqslant d\leqslant D_{port}-1, 0\leqslant p\leqslant P_{proto}-1$，$s,d,p$分别定义为源端口、目的端口、协议字段的范围ID值，S_{port}，D_{port}，P_{proto}分别代表各个字段的分段数。设有如下规则(如表3所列)，则Hash函数处理如图5所示。

表3 规则集

clsID	Sport	Dport	protocol
1	0～30	0～50	6
2	30～80	0～50	17
3	80～65535	50～65535	255

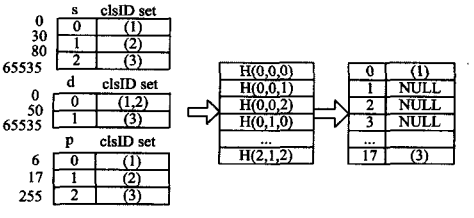

图5 Hash过程

根据式(1)，类似可得IP地址的无冲突Hash函数[17]：

$$H(a,b)=b\times A_{src}+a \quad (2)$$

$$H(c,d)=d\times Index_{A\_table}+c \quad (3)$$

其中a,b表示源/目的IP地址的范围ID值，代表源IP地址的分段个数，c,d表示A_table，P_table的某项索引值，$In-$

dex_{A_table}代表A_table的索引个数。

3.2.3 性能分析

通常分类器的分类字段主要集中于IP帧相关字段上，经过IP分类后规则的个数大大缩小，因此BH算法的时间和空间复杂度主要集中在IP分类结构之上。假设有n个规则，m阶B树共有p个结点，k个端点(关键码)，其深度为h，则有$k\leqslant 2n$。根据B树的定义，可知：

$$h\leqslant \log_{\lceil m/2\rceil}\left(\frac{k+1}{2}\right)+1\leqslant \log_{\lceil m/2\rceil}\left(n+\frac{1}{2}\right)+1 \qquad (4)$$

IP分类有5个字段，需要查找5棵B树，然后进行2次Hash表查找，则IP分类的时间复杂度近似为：

$$O(5(\log_{\lceil m/2\rceil}\left(n+\frac{1}{2}\right)+1)+2)$$

即$\log_{\lceil m/2\rceil}(n)$。当$m$取值越大时，时间复杂度越低。

其中算法在更新时可能会修改多个结点，并在最差情况下会分裂或者合并结点。由于根结点至少有一个端点，其他各非失败结点至少有$\lceil m/2\rceil-1$个端点，则B树至少有$1+(\lceil m/2\rceil-1)(p-1)$个端点，则平均分裂结点数$S$为：

$$S=\frac{\text{分类结点总次数}}{k}\leqslant\frac{p-2}{1+(\lceil m/2\rceil-1)(p-1)}<\frac{1}{\lceil m/2\rceil-1} \qquad (5)$$

当有n个规则时，在最坏的情况下，每个B树端点最多为$2n$，表A_table最多有$2n\times 2n$个表项，表P_table最多有$2n\times 2n\times 2n$个，表AP_table则有$O(n^5)$。与RC-FST性能的对比如表4所列，通过改变m的取值可以对算法的时间空间性能进行调整。

表4 性能比较($d=5$)

	时间复杂度	空间复杂度
RC-FST	log(n)	n^5
BH	$\log_{\lceil m/2\rceil}(n)$	n^5

结束语 本文通过分析和对比常用报文分类算法的性能，针对HFC宽带网络接入技术的CableModem系统QoS体系提出了一种BH分类算法。分析表明，该算法具有时间和空间复杂度小的优点，满足实际嵌入式开发的要求，并且通过控制B树的参数，能对算法的时间空间性能进行调整，从而满足不同应用的需求。本算法对于宽带无线接入网802.16协议的QoS分类器、无线局域网802.11e、路由器等网络协议QoS分类算法的软/硬件实现具有参考借鉴意义。本研究的下一步工作将对BH算法结构进行优化，进一步提高其查找和更新性能。

参考文献

[1] 林锐，单志广，任丰原，等. 计算机网络的服务质量. 北京：清华大学出版社，2004

[2] Yu Lei, Deng Ya-ping, Wang Jiang-bo, et al. A Novel IP Packet Classification Algorithm Based on Hierarchical Intelligent Cuttings// The 6th International Conference on ITS Telecommunication Proceedings. 2006: 1033-1036

[3] 田珂，朱清新，向培素. 一种改进的多维高速报文分类算法. 计算机应用研究，2007(2): 27-32

[4] Hsu Chia-ren, Chen Chien, Lin Chun-Yuan. Fast Packet Classification Using Bit Compression. IEEE Globecom, 2005: 739-743

[5] Abdelghani M, Sezer S, Garcia E, et al. Packet Classification Using Adaptive Rules Cutting// Proceedings of the Advanced Industrial Conference on Telecommunications. 2005

[6] 汪伟，孙翌. 报文分类算法的设计与实现. 上海电力学院学报，2006, 22(1): 63-70

[7] Yu Lei, Deng Ya-ping, Wang Jiang-bo, et al. A Novel IP Packet Classification Algorithm Based on CrossProduct and Hash Tree // The 6th International Conference on ITS Telecommunication Proceedings. 2006: 1037-1040

[8] Tan Xing-ye, Zhan Yong, Leit Zhen-ming. A New Fast Packet Classification Algorithm: RC-FST. IEEE, 2005: 462-466

[9] Zheng Kai, Liang Zhiyong, Ge Yi. Parallel Packet Classification via Policy Table Pre-Partitioning. IEEE Globecom, 2005: 73-78

[10] Sun Xuehong, Sahni S K, Zhao Yiqiang Q. Packet Classification Consuming Small Amount of Memory. IEEE/ACM Transacton on Networking, 2005, 13(5): 1135-1145

[11] Xu Zhen, Sun Jun, Zhang Jun. A Novel Hash-based Packet Classification Algorithm. ICICS, 2005: 1054-1059

[12] Taylor D E, Turner J S. ClassBench: A Packet Classification Benchmark. IEEE/ACM Transactions on Networking, 2007, 15(3): 499-511

[13] Jelassi O, Paul O. Markers - based Space Decomposition Algorithm: A new algorithm for multi-fields packet classification. IEEE, 2006: 43-47

[14] Lu Haibin, Sahni S. O(logW) Multidimensional Packet Classification. IEEE/ACM Transactions on Networking, 2007, 15(2): 432-442

[15] Cable Television Laboratories, Inc. Data-Over-Cable Service Interface Specifications DOCSIS 1.1 Radio Frequency Interface Specification CM-SP-RFIv1.1-C01-050907[S]. 2005

[16] Nu K, Wu J P, Yu Z C, et al. A Non-Collision Hash Trie-ree Based Fast IP Classification Algorithm. J. Computer Sci. & Technol., 2002, 17(2): 219-226

[17] 刘惠义，董志勇，秦益，等. 基于无冲突哈希Trie树的IP分类算法的研究. 计算机与现代化，2004(5)

(上接第121页)

[5] Bellovin S, Merritt M. Limitations of the Kerberos Authentication System. Computer Communications Review, October 1990

[6] Meyer C, Matyas S. Cryptography: A New Dimension in Computer Data Security. New York: Wiley, 1982

[7] Kohl J. The Use of Encryption in Kerberos for Network Authentication// Proceedings, Crypto'89. New York: Springer-Verlag, 1989

[8] Kohl J, Neuman B, Ts'o T. The Evolution of the Kerberos Authentication Service. IEEE Computer Society Press, 1994

[9] Kohl J, Neuman C. The Kerberos Network Authentication Service (V5). Digital Equipment Corp, 1993

[10] Stallings W. 密码编码学与网络安全——原理与实践. 第3版. 刘玉珍，等译. 电子工业出版社，2004

[11] Moron. Guide to Kerberos. 1996. http://www.isi.edu/gost/brian/security/kerberos.html

[12] Butle F, Cervesato I, Jaggard A D, et al. Formal Analysis of Kerberos 5. Theoretical Computer Science, 2006, 367: 57-87

[13] Cervesato I, Jaggard A D, Tsay Joe-Kai, et al. Breaking and Fixing Public-Key Kerberos. Information & Computation, 2007