最大序列匹配问题及路由器解决方法

Maxium Sequence Matching and the Solution Based on Ronter

段 鲲 戴方虎 吴时霖

(复旦大学计算机系 C&C 实验室 上海200433)

Abstract This paper introduces the limitation of IP addresses and B class net. Though CIDR Solves this problem in some degree, it brings in the maxium sequence matching, which is time-consuming. IP swithcing and tag switching are two typical solutions. The analysis and comparation are presented.

Keywords CIDR protocol, Maxium sequence matching, IP switching, Tag switching

随着 Internet 用户数量的激增,突破 IP 路由器的性能限制已成为当前研究的热点,过去,路由器每个端口的价格要远高于交换机端口的价格。随着对路由器研究的深入,这一情况也相应改变。传统的路由器采用的是集中式控制结构,通常用一到两个处理器来实现其功能。为了增加通信量,最新的路由器采用了与交换机相类似的分布式体系结构[1,2],如图1所示。其中,中心控制处理器仍执行路由协议(如,BGP,OSPF 等)和管理协议(如 SNMP,ICMP 等),不过处理和转发 IP 分组的任务已经由输入/输出端口处理器来完成。因此,中心控制处理器创建分组转发表(packet forwarding table),输入/输出端口处理器要复制此分组转发表。中心控制处理器负责在路由发生变化时对分组转发表进行更新。它的基本路由方法与传统的路由器一样,也是根据到来的分组的地址,以一定算法在分组转发表中进行从高位到低的匹配,来找到转发的目的地址。

根据 IPv4,按网络大小分,互联网上可有的三种网络:A 类(主机地址24位),B 类(主机地址16位)和 C 类(主机地址8位)。随着 IP 用户数以及互联网上网络数的急剧增长,原有的 IP 地址难以跟上互联网的发展速度。一个突出的问题就是原有的网络分类方法大量地浪费了网络地址。地址的浪费主要是由 B 类网络造成。因为对于大多数的企业或组织来说,拥有16,000,000个地址的 A 类网络过大,而只拥有256个地址的 C 类地址又过小,所以拥有65,536个地址的 B 类网络就成为首选。但实际上 B 类网络对大多数组织来说,还是过大,这种情况一方面造成了 B 类网络数的短缺,另一方面还造成了 IP 地址的浪费和不足。

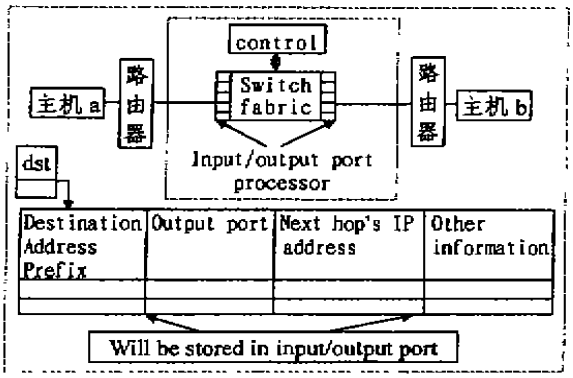


图1 (虚框内为路由器基本结构)

1 CIDR 协议及最大序列匹配问题

针对这一问题,提出了很多解决方法,其中一个正在使用的解决方案就是 CIDR(Classless InterDomain Routing,无类域间路由,由 RFC1519描述),它的基本思想就是以可变的网络大小来重新管理剩余的 C 类网络地址,也就是说主机数地址是以位(bit)而不是以字节(byte)为单位来确定。例如,如果一个网络需要2000个地址,那么就给它分配一个地址数为2048的网络,而不是一个有65,535个地址的 B 类网络。RFC1519还定义了 C 类地址的分配方法。按照此方法,世界被分为欧洲、北美、中南美和亚太四个地区。例如,亚太地区地址的范围为202.0.0.0到203.255.255.255,欧洲地区地址的范围为194.0.0.0到195.255.255.255,这样,分配给每个区域三千二百万个地址,而从204.0.0.0到223.255.255.255的 C 类地址保留为将来使用[3]。

段 鲲 硕士生,研究方向为计算机通信与网络技术。戴方虎 硕士生,研究方向为计算机通信与网络技术。吴时霖 博士生导师,研究方向为计算机通信、网络技术与 Petri 网理论。

现举例如下，假设欧洲有三个组织分别申请地址数为1024，2048，4096的网络。根据 CIDR，可对网络地址进行如下分配：对于所需网络地址数为2048的网络，分配给它的地址范围为从194.24.0.0到194.24.7.255地址掩码为255.255.248.0；对于需要4096个地址的网络，分配给它的地址是从194.24.16.0到194.24.31.255，掩码为255.255.240.0；对于需要1024个地址的网络，其地址为从194.24.8.0到194.24.11.255，掩码为255.255.255.0。对它们的基地址和掩码以二进制分别相应地表示如下：

基地址

11000010 00011000 00001000 00000000
11000010 00011000 00010000 00000000
11000010 00011000 00000000 00000000

掩码

11111111 11111111 11111100 00000000
11111111 11111111 11111000 00000000
11111111 11111111 11110000 00000000

CIDR 虽然在一定程度上缓解了传统的 IPv4所造成的 B 类网络的短缺和 IP 地址的浪费，但对路由器来讲，它引入了新的问题。由于传统的 IP 地址以字节(byte)为单位来定义地址中的网络地址与主机地址，这样在路由器对分组转发表进行查找和匹配时也可按字节进行，从而有较高的速度。如前所述，CIDR 为了将 C 类网络进行合并，是以位(bit)为单位来定义网络地址和主机地址。在路由器根据到来的分组，对分组转发表进行匹配时，就产生了前缀匹配(prefixing match)问题。

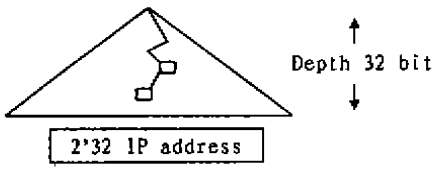


图2

以图2为例，假设采用 CIDR 技术后，路由器的分组转发表中包含有如下两个目的地址前缀：e1＝0011，e2＝001101。此时，路由器收到一个目的地址为0011011011的分组，由于 e2比 e1有更大的匹配位数，故应将其转发到 e2去。在具体实现时，采用选择有最大匹配位数的目的地址作为转发地址，也就是最大序列匹配。但由于需要按位进行匹配，最大序列匹配极为费时。传统的数据结构如哈希表或 Patricia 树难以胜任最大序列匹配[4,5]。即使最新提出的快速算法，配以高速的处理器仍不能很好地在软件上解决这一问题。而采用诸如利用较大的存储器来存储路由表(地址局域性原理)的方法也因为网络用户的增加变得比较困难。根据1988年的研究，只要把20条表项存入快存就可达到超过90%的匹配率[6]，由于网络用户的数量的飞速增长，在1996年的研究表明，要达到相同的性能须存入5000条表项[7]，尤其对于处在骨干网上的路由器，即使是低速的调制解调器，由于其巨大的数目，也会产生很大的流量，所以硬件方法也不能很好地解决这一问题。

2 对最大序列匹配问题的解决方法

解决这个问题的一个途径就是将 ATM(异步传输模式)的一些概念引入到 IP 路由器中，使 IP 在传输时建立类似于 ATM 中的 VP(虚通道)、VC(虚通路)这样的路径。两种典型的方法是 IP 交换和标签交换。下面对这两种方法进行介绍。

2.1 IP 交换

IP 交换[8,9]将 ATM 的交换硬件同传统的 IP 路由器相结合，以提高其性能(见图3)。图3中的中心控制器除实现图1中的中心控制器的功能外，还具有建立连接的能力。IP 交换将连接标识为流(flow)，即所建立的一个端到端的连接。这样的连接通过一个简化了的信令协议来建立，建立以后会对流分配一个虚通路号(virtual channel number)，并把它存储到输入端口处理器。流建立好以后，流上的数据就可以象在 ATM 上那样快速方便地实现交换。显然，当一条路由发生变化时，所有使用此路由的流都得重新进行连接建立的工作，路由器中相关的旧的路由信息将被更新。因此，通常称 IP 变换的连接状态为"软状态"，而传统的 ATM 交换为"硬状态"。流标识根据 IP/TCP/UDP 分组头中诸如服务类型、协议、源/目的端口等信息来确定。而这也就意味着 IP 交换的中心处理器要查看到来分组的第四层的信息(传输层)。

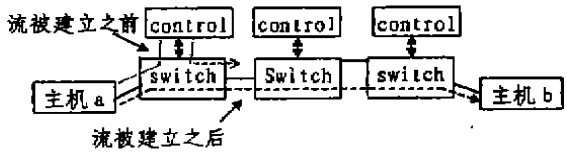


图3

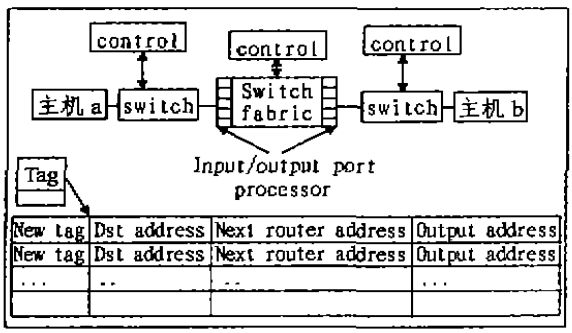
几个潜在的因素有可能会限制 IP 交换的性能。首先，在一个流量较大的网络上，活跃流(active flow)的数量可能会很大(如对于 gigabit 网，可达到十万这样的数量级)。由于交换机内部表容量的限制，所要建立的活跃流只有一小部分可以建立。其次，对于一个建立的连接，很可能它只有很少的信息量(如 telnet 会话)，这就造成了带宽的浪费。另外，由于一个路由的改变会造成多个连接的重新建立，当一次要建立的连接数过

多时，就很可能限制 IP 交换的性能。

2.2 标签交换

标签交换[10]是解决传统路由器中最大序列问题的另一方案。同 IP 交换相似，标签交换也将目的地址转化为一个较短的交换标识。与 IP 交换相区别的是，IP 交换对特定的会话(session)建立通道，而标签交换对路由器中的路由建立通道。从概念上讲，标签交换与 ATM 中的虚通道建立相类似，只不过标签交换要根据路由协议动态地更改这些通路，而对于 ATM，一条虚通路在其存在的过程中通常是不变的。

在路由协议确定路由以后，标签交换将目的地址映射为标签。对于 ATM，其相应的过程为确定虚通路标志和虚通道标志。这个过程被称为标签绑定过程。标签绑定过程需要交换映射信息，这也同 ATM 在连接建立时交换 VC 信息相类似。绑定工作完成以后，每个 IP 分组会附有一个标签，路由器将此标签用于检索输出端口信息，其过程与 ATM 从 VC 翻译表获得输出端口信息相似。



New tag	Dst address	Next router address	Output address
New tag	Dst address	Next router address	Output address
...	..	..	...

图4

2.3 IP 交换与标签交换的比较

IP 交换与标签交换的根本区别就是会话(即流)和路由的区别。一个路由可包含多个会话，因此在一个链路所承载的路由数量要远小于上面的会话的数量。所以，同样条件下当路由发生变化时，IP 交换所要重建的连接数量要远大于标签交换要重建的数量。另外，标签交换不需要查看传输层协议。同时 IP 交换和标签交换存在一个根本的相似点：它们都没有遗弃传统的 IP 处理。对于 IP 交换来说，路由器的中心处理器仍要处理非流数据(如在连接建立阶段)；对标签交换来说，其网络的边缘路由器还要和其它的非标签交换的路由器交换数据。

总结 本文介绍了路由器的基本构造，并简介了在当前网络和网络用户迅速增长条件下，CIDR 使 IP 地址和 B 类网络不足问题得到一定程度上的解决。同时对 CIDR 引出的前缀匹配问题进行了分析。对于此问题，从路由器角度讲，常见的有两种解决方法：IP 交换和标签交换。两者都是在传统路由器基础上引入了一些 ATM 的思想，实际上就是在对无连接的服务具有面向连接的特性，用特定的标识来取代 IP 中的地址从而解决了路由过程中地址匹配过于耗时的问题。

参 考 文 献

1 Tantawy A, et al. On the Design of Multi-gigabit IP Router J. High Speed Networds, 1998, 3(June): 209～232

2 Patridge C, et al A 50-Gb/s IP Router. IEEE Trans. Networking, 1989, 6(3): 237～248

3 Tanenbaum A S. Computer Network. 北京：清华大学出版社，1996

4 Leffle S J, et al. The Design and Implementation of the 4.3 BSD UNIX Operating System, Reading, MA. Addison-Wesley, 1989

5 Wright G R, Stevens W R. TCP/IP illustrated, vol. 2, Reading MA. Addison-Wesley, 1995

6 Feldmeier D C. Improving Gateway Performance with a Routing-Table Cache. Proc. IEEE INFOCOM'88

7 Partridge C. Locality and Route Caches. NSF Wksp Internet Statistics Measurement and Analysis. San Diego, CA, Feb 1996

8 Newman P, Lyon T, Minshall G. Flow Labelled IP: A connectionless approach to ATM. Proc. IEEE INFOCOM'9: 1251～1260

9 Parulkar G, Schmidt D C, Turner J. altPm: a strategy for integrated IP with ATM. Proc. SIGCOMM'95, Cambridge, MA: 49～56

10 Rekhter Y, et al. Cisco Systems' Tag Switching Architecture Overview IETF RFC 2150, Feb. 1997