宽带IP路由器的体系结构分析*

徐恪 熊勇强 吴建平

(清华大学计算机科学与技术系 北京 100084)

E-mail: xuke@mail.cs.tsinghua.edu.cn

摘要 随着宽带技术的不断发展,组建主干网的路由器必然需要以千兆比特以上的速率转发分组,而基于总线和中央处理器的路由器具有无法克服的局限,这就对传统的路由器体系结构提出了严峻的挑战.该文全面综述了近年来在宽带IP(Internet protocol)路由器方面研究的最新进展,详细分析了用于主干网互连的宽带IP路由器的体系结构设计,最后,指出了该领域中需要进一步研究的问题.

关键词 宽带,IP(Internet protocol),路由器,体系结构.

中图法分类号 TP393

随着Internet的飞速发展和宽带技术的不断出现,对Internet互连的核心设备——路由器的性能提出的要求越来越高.传统的基于总线和中央处理器结构的路由器由于其体系结构上的局限,已经无法满足组建高速主干网络的需求.近年来,国际上对宽带IP路由器技术的研究也日益活跃,提出了几种不同的研究和设计思路[1,2],其基本思想主要包括4个方面:(1)将路由引擎(routing engine)和转发引擎(forwarding engine)分开,将局部转发表从全局路由表中独立出来;(2)用快速的硬件实现IP报文的报头处理、寻径和转发;(3)用多个分布式的接口单元加中央控制器的模式取代中央处理器加接口卡的模式;(4)用交换结构(switch fabric)提高各接口单元之间的数据通信速度.从研究趋势来看,IP路由技术的发展已经和交换技术以及宽带技术的发展有机地结合在一起.本文综述了目前宽带IP路由器所采用的主要的体系结构,详细讨论了其中的技术难点和相应的解决方案,并提出了进一步的发展方向.

本文第1节分析IP路由器的4种主要体系结构.第2节论述宽带IP路由器设计中存在的主要问题.第3节讨论了当前解决这些问题的主要的方案.第4节总结了全文,并指出了进一步的研究方向.

1 IP路由器的体系结构分析

从体系结构来看,IP路由器经历了从单处理器到并行处理器、从共享总线到交换结构的发展过程.我们可以把它划分为以下4种类型.

单处理器共享总线式体系结构.这是第1代路由器主要采用的体系结构:基于单个通用CPU,使用实时操作系统(如图1(a)所示).采用这种体系结构主要是考虑到网络协议经常发生变化,而运行多个协议的路由器不可能针对某种特定的协议进行优化,此时连接的建立和管理比高转发能力更重要.这种体系结构的路由器可以使用通用计算机来实现,与一般的通用计算机不同的是,它具有多块网络接口卡,网络接口卡之间通过系统总线相连.到达网络接口的报文首先被送到中央处理器,由中央处理器上运行的路由引擎决定下一跳的地址,并把它送到相应的输出网络接口上.路由协议和其他控制协议均在中央处理器上实现.

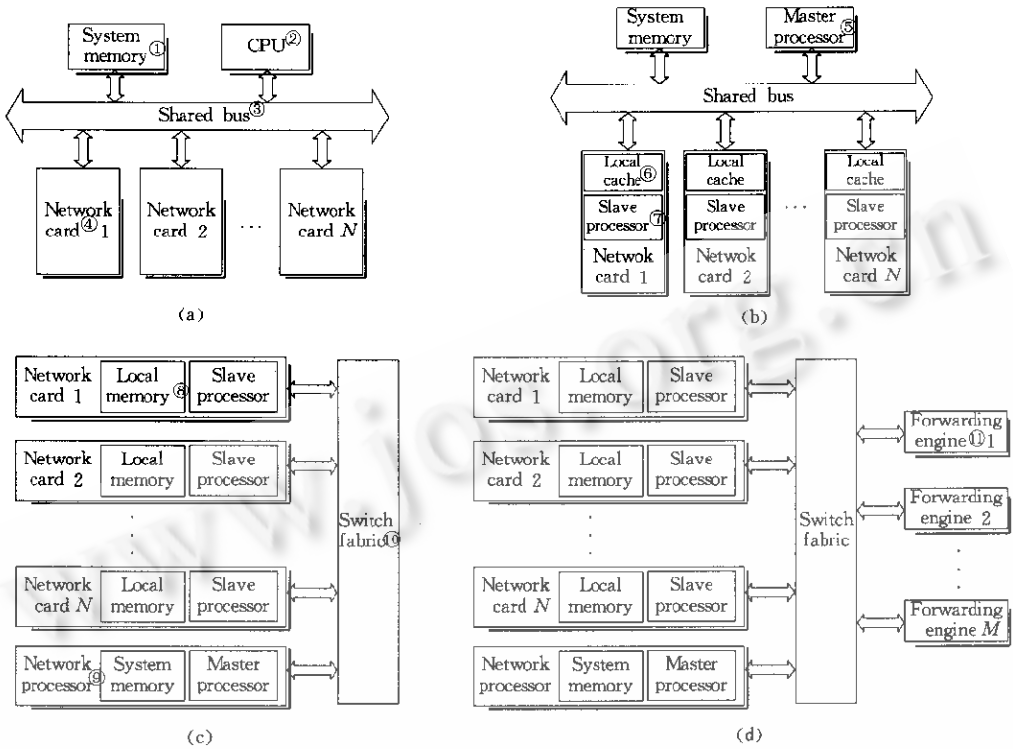
显然,这种路由器的性能主要由共享总线的吞吐率和主CPU转发报文的速度决定.由于主CPU必须执行

* 本文研究得到国家自然科学基金(No.69582002,69725003)和国家"九五"科技攻关项目基金资助.作者**徐恪**,1974年生,博士生,主要研究领域为计算机网络体系结构,计算机系统性能评价.**熊勇强**,1974年生,博士生,主要研究领域为计算机网络体系结构,网络安全技术.**吴建平**,1953年生,博士,教授,博士生导师,主要研究领域为计算机网络体系结构,网络协议测试.

本文通讯联系人:徐恪,北京 100084,清华大学计算机科学与技术系网络研究所

本文1999-07-01收到原稿,1999-09-28收到修改稿

多个实时操作,因此,操作系统的选择相当重要,而实时操作系统的设计也比较复杂,因此,这种体系结构的可扩展性(scalability)比较差,而且很难与网络接口卡接口速率的提高相适应.



①系统内存,②处理器,③共享总线,④网络接口卡,⑤主处理器,⑥本地缓存
⑦从处理器,⑧本地内存,⑨路由协议处理器,⑩交换结构,⑪转发引擎.

Fig. 1　The architectures of four kinds of principal routers
图1　4种典型的路由器体系结构

多处理器共享总线式体系结构.在第1代路由器中,所有的路由计算都在中央处理器上进行,在高速和动态变化的网络环境下,路由计算的能力将制约路由器的转发速率.第2代路由器(如图1(b)所示)采用的多处理器共享总线式体系结构的路由器把转发计算分布在各个处理器上,从而有效地解决了路由计算能力的问题.

在多处理器共享总线式体系结构中,网络接口卡具有本地的快速处理器和高速缓存以及具有独立处理分组的能力:每个连接的第1个包被送到主CPU的路由引擎上进行处理,路由引擎在得到输出接口卡的端口号之后,将其传给输入网络接口卡,输入网络接口卡就在本地高速缓存中增加一个表项.这样,该连接以后的分组就可以直接在网络接口卡之间交换而无需再经过主CPU.在这种结构中,路由计算就转换成各分布式处理器上的转发计算了,因此,转发引擎所使用的高速缓存表项的设计就必须比较精巧.我们可以为每个连接建立一项转发表项(one forwarding item per connection,简称OFIPC),也可以每条路由建立一项(one forwarding item per route,简称OFIPR),如果采用OFIPR,即使连接数非常多,高速缓存表项也不会太大.

这种体系结构的主要问题是共享总线的容量限制,共享总线的容量直接限制了路由器的吞吐率,成为系统无法避免的瓶颈.另一个问题是,这种体系结构很难用在主干网络路由器上.由于主干网络路由器往往具有很高的包转发速度,因此,在网络接口卡上的路由一般不具有局部性.在这种情况下,网络接口卡上的高速缓存很难发挥作用,也就很难减轻主CPU的负担.为了解决这一问题,可以在每个网络接口卡上都存放完整的路由表,这样可以进一步增强这种路由器的能力.

多处理器交换式体系结构.为了解决第2代路由器中的系统总线瓶颈问题,人们提出了多处理器交换式体系结构(如图1(c)所示).在这种体系结构中,第2类结构中的系统总线被交换结构所代替.交换结构可以提供比

共享总线高得多的带宽,足以支持现有的高速网络接口.

旧的问题解决了,新的问题又随之而来.在这种路由器中,每个报文的处理成了新的瓶颈.为了提高报文处理的速度,人们又提出了新的体系结构.

共享并行处理器交换式体系结构.在文献[1]中提出了并行报文处理的方法,使用这种方法可以极大地加快报文处理速度.该方法的基本设计思想基于如下考虑,一般来说不可能出现所有的网络接口同时阻塞的情况,因此,可以通过共享转发引擎来提高路由器的端口密度.这种体系结构如图1(d)所示.

在如图1(d)所示的体系结构中,转发引擎只负责查找下一地址,在找到下一地址后,将其送给网络接口卡,这样,网络接口卡就可以直接把后继的报文送交输出接口.需要注意的是,这里只需要把报文的头部送交转发引擎.这样做可以减少互连结构上不必要的报文传输负载.报文体只在接口卡之间传递.

在这种体系结构中,只要有足够的转发引擎,就可以使路由器具备高速主干网络需要的转发能力.使用这种体系结构是因为单一的转发引擎没有足够的能力处理所有的网络接口卡的突发性的传输请求.因此需要在多个网络接口卡之间共享多个转发引擎.

从物理上来说,转发引擎既可以和网络接口卡做在一起,也可以相互独立.如果转发引擎在物理上是独立的部分,那么通过调整转发引擎和网络接口卡的比例就可以增大或者减少包转发率,文献[3]中讨论的50Gb的IP路由器就使用了这种方案.但是,把转发引擎和网络接口卡分开会给交换结构带来额外的负担,而且在多个转发引擎之间还需要使用负载平衡机制.因此,也有的设计方案把网络接口卡和转发引擎集成在一起,如NetStar公司的GigaRouter[4].在这种方案中,转发引擎不在所有的网络接口卡之间共享,而是每个网络接口卡都具有一个转发引擎.实际上,这种体系结构就是多处理机交换式体系结构.使用这种体系结构必须保证单个的转发引擎能够满足网络接口卡的突发性的传送请求.

2 宽带IP路由器设计中的主要问题

目前的宽带IP路由器主要采用第1节中提到的第3种和第4种体系结构.下面,我们将主要讨论这两种体系结构的路由器的设计问题.

在宽带IP路由器中,可能的系统瓶颈有两处:一处是报文的转发.在共享总线式的结构中,这是系统的主要瓶颈.在交换式结构中,由于交换结构的带宽很高,报文转发不再是系统的主要瓶颈.另一处是报文的路由,即在转发表(也称为路由表)中查找路由.当一个报文到达时,路由器必须根据其目的地址在转发表中查找目的端口号.一般来说,转发表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〉;当接受到目的地址为A的报文时,输入端口将会遍历转发表,使用掩码和A进行与操作,如果结果和相应的网络地址一样,该端口就成为候选端口之一.最后选择的端口将是所有候选端口中相应的掩码最长的端口,这被称为最长前缀匹配(the longest prefix match).举例来说,如果一台路由器的路由表中有如下3项:〈166.111.1.5/16,1〉,〈166.111.225.0/18,3〉,〈128.0.0.0/8,5〉.那么,一个目的地址为166.111.195.1的报文将会同时匹配前两项.由于端口3对应的路由表项的掩码最长,因此,该报文的目的端口应该是端口3.

从这个例子中可以看出,高效的路由表查找是相当困难的.首先,主干路由器上路由表包括上万条表项(一个典型的值是15K路由表项),如果顺序查找,其效率将非常低.为了尽可能提高路由速度,我们必须尽可能减少路由查找的时间.其次,一个报文可能会同时匹配多个路由表项,我们必须在这多个路由表项中寻找最长的匹配.前面也提到过,在主干网络上,使用路由表高速缓存并不会提高路由表查找的速度.因此,我们还需要使用新的方法来提高路由表的查找速度.

根据输入/输出端口和交换结构的相对速度的不同,宽带IP路由器可以分成输入排队的(input-queued)和输出排队的(output-queued)路由器.如果交换结构的带宽大于所有的输入端口的带宽的总和,那么报文就只在输出端口排队,这种路由器被称为输出排队的路由器.否则,报文就会在输入排队,这种路由器就被称为输入排队的路由器.

输入排队的路由器和输出排队的路由器都有路由查找的瓶颈,除此之外,它们各自还具有自己特有的性能瓶颈.由于在输出排队的路由器中所有的报文都在输出端口的缓冲区排队,而输出端口缓冲区的存取速度是受

动态存储器和静态存储器的访问时间限制的.这最终将会限制输出排队路由器的性能.那么,我们如果把所有的队列都放到输入一端,又如何呢?如果这样做,就需要由仲裁器来解决交换结构和输出线之间的竞争问题,但是设计一个高速而公平的仲裁器是相当困难的[5,6].

3 主要问题的解决方案

本节将结合近年来在宽带IP路由器研究和设计中取得的新进展来阐述如何解决第2节中提到的几个主要问题.

3.1 高效的路由表查找

在第2节中我们已经提到,在使用交换式结构的宽带IP路由器中,主要的性能瓶颈就是在路由表中查找与到达报文的目的地址相匹配的最长前缀的路由.我们知道,路由表查找的速度是由查找时需要访问内存的次数和内存的速度共同决定的.例如,一种算法需要执行8次内存访问,而每次内存访问的时间是60ns,那么查找一次路由的时间就是480ns,也就是说每秒可以执行大约200万次路由查找.使用相同的算法,如果使用访问速度为10ns的内存,那么每秒就大约可以执行1 250万次路由查找.

在设计路由表查找算法时需要考虑的另一个问题是路由表的修改.最近的研究表明,路由表的变化是比较慢的,大约只需要每两分钟更新一次[7],这就允许我们使用比较复杂的数据结构来提高查找速度,而付出的代价是修改路由表需要更多的时间.

常见的保存路由表的数据结构是树,在树中,每条从根节点到叶子节点的路径就对应着路由表中的一项.这样,寻找最长的前缀就转化成为寻找最长的路径[8].一般来说,基于树的算法从树的根节点开始,使用目的地址中的下若干位来匹配当前节点的子节点,直到找到一个匹配为止.因此,在最坏情况下查找路由表所花费的时间和找到的最长前缀匹配的长度成正比.基于树的算法的主要思想是大多数节点只需要保存很少的子节点而不用保存所有可能的值.这类算法节约了内存,付出的代价是需要进行更多次的内存查找.随着内存价格的下降,这种设计方法已经越来越不常用了.

提高路由表查找性能的方法有多种,我们可以把它们分成以下3类:

(1) 基于硬件的技术

常用的基于硬件的技术使用相联存储器(content-addressable memories)和高速缓存来提高查找速度.这两种技术都很难随着路由表的变化而扩展,因此不能用在需要使用大路由表的主干路由器上.近来,研究人员提出的基于硬件的策略把内存和逻辑集成到一块芯片中,这样可以极大地减少内存访问时间.

第2种基于硬件的技术是通过增大使用的内存来存储路由表,通过空间换取时间.文献[9]提出了一种方案,使用33M的内存来保存路由表.达到的性能是在多数情况下,只需要查找一次路由表就可以得到目的端口,在最坏的情况下,也只需要查找两次路由表.通过使用一些优化技术,可以把内存的使用量降到9M,相应的最坏查找时间就会变成3次内存访问时间.该方案的基本思想是:把32位的目的地址分成两部分,前24位是一部分,后8位是另一部分.使用前24位作为地址直接查找路由表,如果目的地址的路由长度小于24位,则会在一次内存访问时间内查找到目的端口.如果目的地址的路由长度大于24位,则在路由表的相应位置会得到一个索引值,利用该索引值和目的地址的最后8位作为地址去查找另一个存储器,这样,在最坏的情况下,也只需要读两次内存.该方案的核心思想就是把路由表分散保存,例如,1条(166.111.0.0/16,1)的路由将在路由表中占据从地址166,111,0,0开始的连续的256项.这样做虽然提高了路由表的查找速度,但却带来了路由表更新的不便.1条路由的改变就需要更新大量的路由表项.于是,为了提高更新的速度,在文献[9]中提出了相应的解决方案.

(2) 表紧缩技术

表紧缩技术[10]是使用复杂但是紧缩的数据结构来保存路由表.这样,路由表就可以放在处理器的第1级高速缓存中.这种方案可以支持千兆比特速率的路由表查找.使用文献[10]中提出的数据结构,可以把具有40 000条路由的路由表压缩到只有150K字节.

(3) 哈希表技术

哈希表技术也可以用于路由表查找.寻找最长前缀匹配的需求限制了哈希表技术的使用.在实际使用中,我们并不知道一个目的地址对应的最长前缀匹配是多长.解决这一问题的方法是尝试不同长度的掩码,从中选择最长的匹配.掩码的选择既可以使用迭代方式,也可以使用层次方式,还可以使用地址的前几位指向一个前缀长度列表.但是,这些方案的可扩展性都很差.

文献[11]提出了一种可扩展的基于哈希表的查找算法,该算法查找 n 位地址的最长前缀匹配的步骤为 $O(\log n)$.该算法对于每种长度的前缀都使用一个单独的哈希表,查找时并不是从最长的前缀开始,而是根据前缀长度执行折半查找.为了能够执行折半查找,就需要哈希表包含标记字段,标记字段用于指出当查找失败时,是到长度更大的表中查找,还是到长度小一些的表中查找.此外,文献[11]还采用了相应的技术来减少标记字段的存储空间.该算法可以比较容易地扩展到IP v6.

3.2 高速的报文转发

交换结构是宽带IP路由器中的关键部分,是解决高速报文转发的主要方式,它的性能直接决定了路由器的性能.交换结构一般可以采用交叉开关和共享内存两种方式来实现.交叉开关的速度受调度器的限制,而共享内存的速度受内存访问速度的限制.Cisco公司推出的高端路由器GSR12000系列,使用了交叉开关作为交换结构[12],而Juniper公司推出的主干网路由器Juniper M40则使用了共享内存作为交换结构[13].交叉开关和共享内存都能够达到比较高的吞吐率,即都能够达到每秒40Gbps(当配置8块OC-48线路接口卡时)以上的吞吐率.共享内存的特点是实现简单,也能达到比较高的吞吐率,但是其性能的进一步提高将受到内存访问速率的限制,而且其可扩展性比较差,当线路接口卡数量较多时,性能将受到一定的影响.而交叉开关能够达到比较高的速率,扩展性好,但是需要设计完善的调度算法并用高速硬件实现调度器.随着人们对交叉开关调度算法研究的深入,已经设计并实现了许多性能良好、实现简单的调度算法,因此,目前宽带IP路由器都趋向于使用交叉开关作为交换结构,下面我们将重点讨论交叉开关的设计与实现.

使用交叉开关主要基于如下的考虑:首先,交叉开关可以在网络接口卡之间建立点到点的连接,这就意味着网络接口卡之间可以高速传递数据.目前,在商用产品中,芯片到芯片的线路速度已经可以达到1Gb/s,在实验室中已经可以达到4~10Gb/s.其次,交叉开关可以同时提供多路传送,只要源和目的地不冲突,就可以同时传送,这样可以极大地增加带宽.

在使用交叉开关时,需要把报文变成定长的分组.在使用定长的分组时,可以根据分组的大小划分时间片,根据时间片一步一步地处理.如果使用变长的分组,那么分组通过交叉开关的时间就是随机的,调度器就必须知道所有输入和输出的状态,这使调度器的设计相当复杂,而且很难做到公平调度.

在使用交叉开关时,需要解决以下几个主要问题.

3.2.1 阻塞问题

当使用交叉开关作为交换结构时,可能会遇到以下3种阻塞.

第1种是线路头部阻塞.如果输入端口等待交换的分组都使用一个队列排队,就会出现线路头部阻塞问题.举例来说,如果队列头部的分组要去的端口正忙,那么此分组只能在队列中等待.这时即使它后面的分组要去的端口是空闲的,也没有机会发送.这种阻塞会极大地降低交叉开关的流量.在文献[14]中提出了一种称为虚拟输出队列(virtual output queuing,简称VOQ)的机制来解决线路头部阻塞问题.思路是:假设系统中有 n 个输入端口和 n 个输出端口,那么每个输入端口都有 n 个队列分别对应每个输出端口.这样,若一个输出端口发生阻塞对其他的输出端口不会产生影响.当然,在实际输出时,每个端口只对应于交叉开关中的一条线,因此称之为虚拟输出队列.

第2种是输入阻塞.由于虚拟输出队列对应的交叉开关的线只有1条,因此,每次只能交换一个分组.当虚拟输出队列中有多个分组时,得不到交换机会的队列头部分组就处于输入阻塞状态.输入阻塞不影响交叉开关的流量,只会增大被阻塞的分组的延迟.

第3种是输出阻塞.如果两个输入端口的分组都要去同一个输出端口,就会发生输出阻塞.这种阻塞和输入阻塞一样,不会影响交叉开关的流量,只增大被阻塞的分组的延迟以解决输入阻塞和输出阻塞.在文献[12]中提

出了两种解决方案.第1,带优先级的虚拟输出队列.给每个输出队列划分4个优先级,相应的输出队列也就变成4条,高优先级的分组优先发送.这种方案并不能完全解决输入阻塞问题,比如优先级相同的分组之间还是有阻塞,但是它可以保证高优先级的分组的延迟比较小.第2,加快交叉开关的交换速度.如果交叉开关的交换速度是端口速度的两倍,那么,相对于端口来说,交叉开关能够一次交换两个分组.从理论上说,如果有n个输入端口和n个输出端口,那么交叉开关的速度必须是端口速度的n倍才能保证没有输出阻塞.文献[12]指出,在实际使用中,只要交叉开关的速度是端口速度的两倍,就可以基本保证不出现输出阻塞.

3.2.2 调度算法

交叉开关的另一个重要问题就是调度算法.调度算法设计的基本要求是:

(1) 效率高.高效率的调度算法应该能够同时匹配尽可能多的输入队列.一般来说,使用硬件很难快速计算出最佳匹配,因此,一般在设计调度算法时总是尽力寻找次优的算法.

(2) 稳定性.无论输入队列的情况如何,调度算法都应该迅速找到可行的调度.

(3) 不会出现某些队列永远得不到响应的情况.

(4) 快速.调度算法必须快速执行,否则将会抵消交叉开关带来的带宽的增加.

(5) 易于实现.调度算法应该易于用硬件实现.实现的复杂性包括调度器维护的状态的数量,基于这些状态作出决策的逻辑的复杂度和修改状态时的通信开销.

调度算法可以分成输入排队的调度算法和输出排队的调度算法.长期以来,人们一直认为输入排队的调度算法性能比较差,因而对输出排队的调度算法进行了大量的研究[15~17].但是,输出排队的调度算法要求输出端的接口速率是输入端的N倍(N是端口数量),否则就会出现大量丢包的情况.随着输入端口速度的不断提高和输入端口数量的增多,输出排队的调度算法已经不能满足高速交叉开关的要求.因此,研究人员又重新把注意力集中到了输入排队的调度算法上.

在文献[18]中提出了一种输入排队的称为iSLIP(iterative round-robin matching with SLIP)的交叉开关调度算法.该算法是一个简单的循环算法,每次循环建立一个或多个输入到输出的连接为止,直到建立所有可能的连接为止,然后同时传送.每次循环分成3步,第1步是输入端向输出端提出请求;第2步是每个输出端在所有的请求中选择优先级最高的进行回应;第3步是每个输入端从可能的多个回应中选择优先级最高的建立连接.为了保证公平调度,使用了优先级轮转机制.为了能够处理多播传送,文献[18]把iSLIP算法扩展成可以支持多播的ESLIP(extended SLIP)[18].该算法原理简单,易于用硬件实现,而且具有很好的性能[5].Cisco公司推出的高端路由器GSR12000系列中的交叉开关调度算法就采用了iSLIP和ESLIP.

调度算法的进一步的研究方向是提供对服务质量QoS(quality of service)的支持.在文献[19]中提出了一种结合输入、输出排队的交换结构CIOQ(combined input/output queuing),并提出了若干调度算法,该交换结构能够在保证高分组交换率的同时提供服务质量支持,该交换结构的加速比为2.在文献[20]中,对输入排队的交换结构的服务质量支持进行了研究,并提出了几种线性复杂性且不需要加速比的调度算法.如何基于流进行支持服务质量的调度将是调度算法的进一步的研究方向.

3.3 高性能的实时路由器操作系统

以前的路由器操作系统一般都是一个简单的内核,提供基本的服务原语.随着网络协议的不断发展,对路由器操作系统的要求也越来越高,路由器操作系统需要能够支持路由协议的高性能的执行,需要能够允许上层协议软件的动态升级.这就需要一种高性能的、模块化的、可扩展的操作系统来支持.

文献[21]提出了基于plugin的路由器操作系统体系结构,在该体系结构中,上层协议软件可以作为plugin动态加载到操作系统内核中执行,以获得高性能和高可扩展性.文献[21]根据此体系结构构造了一个扩展的集成服务路由器.

这种方式的一个极端就是主动网络(active network),在主动网络中,可以通过网络分组给路由器安装协议软件[22].但是这种方式给整个网络带来了很大的安全隐患,能否取得很好的效果还有待于进一步研究.

4 总 结

随着计算机网络互联规模的不断扩大,对主干网络路由器的性能提出的要求越来越高.本文综述了近年来国内外在宽带IP路由器研究领域取得的新进展.随着网络技术的发展,我们必须致力于研究自主版权的宽带IP路由器.在国家"九五"规划期间,清华大学计算机科学与技术系设计和开发了性能超过Cisco公司7000系列路由器的高性能路由器[23],为进一步的研究打下了良好的基础.在高速宽带IP路由器领域中,需要进一步研究的问题有:

(1) 如何支持流标识.流是在一段时间内从某个源地址到某个目的地址的连续的报文.流可能是一个连续的TCP连接发送的报文,也可能是网络会议中的UDP(user datagram protocol)报文.通过定义流,可以优化资源的使用,比如高速缓存表项.在标识流时需要使用分类算法.而前面讨论的高速路由算法都很难被修改成分类算法.目前还缺乏通用而有效的流描述.这些还都需要进行进一步的研究.

(2) 如何支持资源预留.Internet对资源预留的支持很弱,无论是局域网路由器、广域网路由器还是宽带IP路由器采用的都是尽力传送机制,不支持优先级.随着网络应用的发展,人们对网络的服务质量QoS提出的要求越来越高,为了支持服务质量的要求,需要路由器支持资源预留.支持资源预留首先要解决上面提到的流标识问题,在流标识的基础上,还需要实现灵活而通用的调度和缓冲区管理算法,支持服务质量的交换结构和有效的流隔离机制.

(3) 高效率的路由器操作系统.以前的路由器往往被看成是转发IP分组的硬件设备.因此,以往的路由器操作系统往往功能很少,而且一般都不提供应用编程接口API(application programming interface).随着网络应用的发展,端用户和网络管理员越来越需要动态地往路由器中加载软件模块以提供防火墙、流量管理等机制.这就要求路由器操作系统能够提供一个灵活的、高效率的机制来支持这种应用需求.

参考文献

1 Asthana A, Delhp S, Jagdish H *et al*. Design of a Gigabit IP Router. Technical Report, 11251-911105-09TM, AT&T Bell Laboratorys, 1991

2 Asthana A, Delhp S, Jagdish H *et al*. Towards a gigabit IP router. Journal of High-Speed Networks, 1992,1(4):281~288

3 Partidge C, Carvey P P, Burgess E *et al*. A 50Gb/s IP router. IEEE Transactions on Networking, 1998,6(3):237~248

4 Kachelmeyer D. A New Router Architecture for Tomorrow's Internet. NetStar, Inc., http://www.netstar.com

5 McKeown N. Scheduling algorithms for input-queued cell switchs [Ph.D. Thesis]. Berkeley, CA: University of California, 1995

6 McKeown N, Anantharam V, Walrand J. Achieving 100% throughput in an input-queued switch. In: Toshiharu Hasegawa, Pickholtz R L eds. Proceedings of the IEEE INFOCOM'96. San Francisco, CA: IEEE Computer Society Press, 1996. 296~302

7 Labovitz C, Malan G R, Jahanian F. Internet routing instability. ACM Computer Communication Review, 1997,27(4): 115~126

8 Stevens W R. TCP/IP Illustrated, Vol. 2. MA: Addison-Wesley, 1995. 220~223

9 Gupta P, Lin S, McKeown N. Routing lookups in hardware at memory access speeds. In: Guerin R ed. Proceedings of the IEEE INFOCOM'98. San Francisco, CA: IEEE Computer Society Press, 1998. 1240~1247

10 Brodnik A, Carlsson S, Degermark M *et al*. Small forwarding tables for fast route lookups. ACM Computer Communication Review, 1997,27(4):3~14

11 Waldvogel M, Turner J, Plattner B. Scalable high speed IP routing lookups. ACM Computer Communication Review, 1997,27(4):25~36

12 McKeown N. Fast Switched Backplane for a Gigabit Switched Router. White Paper, http://www.cisco.com

13 Semeria C. Internet Backbone Routers and Evolving Internet Design. White Paper, http://www.juniper.com

14 Mekkittikul A, McKeown N. Practical scheduling algorithm to achieve 100% throughput in input-queued switches. In: Guerin R ed. Proceedings of the IEEE INFOCOM'98. San Francisco, CA: IEEE Computer Society Press, 1998. 792~799

15 Kelly. Effective bandwidths at multiclass queues. Queuing Systems Theory and Applications, 1991,9(1-2):5~15

16 Keshav S. An efficient implementation of fair queuing. Internetworking: Research and Experience, 1991,2(3):157~173

17 Kesidis G, Walrand J, Chang C. Effective bandwidths for multiclass Markov fluidsand other ATM sources. IEEE/ACM Transactions on Networking, 1993,1(4):424~428

18 McKeown N, Izzard M, Mekkittikul A *et al*. Tiny Tera: a packet switch core. IEEE Micro, 1997,17(1):26~33

19 Shang-Tse Chuang *et al*. Matching output queuing with a combined input/output-queued switch. IEEE Journal on Selected Areas in Communications, 1999,17(6):1030~1039

20 Anthony C K, Siu Kai-yeung. Linear-complexity algorithms for QoS support in input-queued switches with no speedup. IEEE Journal on Selected Areas in Communications, 1999,17(6):1040~1056

21 Decasper D, Plattner B, Parulkar G M *et al*. Scalable high-performance active network node. IEEE Network, 1999,13(1):8~19

22 Alexander D S, Shaw M, Nettles S M *et al*. Active bridging. ACM Computer Communication Review, 1997,27(4):101~111

23 Fan Xiao-bo, Lin Chuang, Wu Jian-ping *et al*. Performance model and analysis of a distributed router. Chinese Journal of Computers, 1999,22(11):1~5
(范晓勃,林闯,吴建平等.分布式路由器的性能模型与分析.计算机学报,1999,22(11):1~5)

Analysis of Broadband IP Router Architecture

XU Ke XIONG Yong-qiang WU Jian-ping

(*Department of Computer Science and Technology Tsinghua University Beijing* 100084)

Abstract With the increasing development of broadband technology, the speed of packets forwarding needs to be over gigabits per second for the backbone routers. Traditional routers have some insuperable barriers and can not solve the problem based on shared-bus and central processing unit. It is a great challenge for the future router architecture. In this article, the authors survey the recent advances in the research of broadband IP router, and analyze the architecture design of the fourth generation backbone router in detail. Finally, some challenging open problems are identified.

Key words Broadband, IP (Internet protocol), router, architecture.