IP 路由器体系结构的综合研究*)

Comprehensive Study of IP Router Architecture

刘宴兵　李　春　幸云辉

(重庆邮电学院计算机系　重庆400065)　(北京邮电大学　北京100876)

Abstract　With Internet and 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 great challenge for future router architecture. In this paper, the authors survey the recent advances in the research of broadband IP switch router, and analyze the architecture design of the fourth generation backbone router in detail. Finally, some comprehensive appraising towards router architecture are identified.

Keywords　IP, Router, Broadband, Switch, Architecture

1　引言

交换节点的资源路由调度是保证实时业务服务质量和提高网络资源利用率的关键[1]，从研究趋势看，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。路由器通常把数据包从一个数据链路中继到另一数据链路，为了中转数据包，路由器具有两个基本功能：路径判断和交换。近年来，国际上对宽带 IP 技术上的研究也日益活跃，下面就 IP 路由器体系结构进行综合研究。

2　IP 路由器的体系结构发展历程

从体系结构看，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。

2.1　单处理器共享总线式体系结构

这是第一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。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伸缩性(Scalability)比较差，而且很难与网络接口卡接口速率的提高相适应。

2.2　多处理器共享总线式体系结构

在第一代路由器中，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。第二代路由器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

*)基金项目：重邮青年教师科技基金项目(No. A2001-20)和信息产业部"95"发展项目(No. 98048)。刘宴兵　讲师，硕士；幸云辉　教授。

速缓存表项的设计就必须比较精巧。可以为每个连接建立一项转发表项(On Forwarding Item Per Connection，简称 OFIPC)，也可以每条路由建立一项(On Forwarding Item Per Router，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。

2.3 多处理器交换式体系结构

为了解决第一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。在这种体系结构中，第二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。

2.4 共享并行处理器交换式体系结构

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。转发引擎只负责查找下一地址，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。

3 交换式 IP 路由器体系结构特点剖析

交换式路由器采用分布式的体系结构，一个路由器内有多个处理单元，可以同时对进入线路卡的分组进行寻路。总的来说，这种交换式路由器具有以下特点：

(1)只支持 TCP/IP 协议。随着网络的发展，TCP/IP 协议已经在网络上占据了统治地位。交换式路由器只需要支持 TCP/IP 协议族，这样就大大减轻了路由器的负担，加快了路由器处理分组的速度。

(2)数据通道与管理通道部分分开。对每一个分组，都需要根据分组的目的地址进行寻路。这部分操作对每个分组都需要执行，必须加以优化以提高处理速度。相对于数据通道，运行路由协议以及进行网管等操作的实时性要求不高，这部分可以和数据通道分开。CPU 卡上的 CPU 运行路由协议，生成路由转发表，并且将完整的路由转发表下载到所有线路接口卡上。线路端口收到分组后，根据接口卡上的路由转发表对分组寻路，然后将分组送往目的输出端口。输出端口收到分组后将分组送往物理线路。由于线路卡上的操作相对简单，可以专门加以优化，从而提高了路由器处理分组的速度。

(3)采用交换结构来代替共享总线方式，从而大大提高了系统的总容量，也由于这个原因，称之为交换式路由器。

(4)许多路由器能够根据 IP 分组的源 IP 地址、目的 IP 地址、协议类型、源端口号和目的端口号对分组进行分类，对不同种类的分组提供不同的服务。这种处理分组的方式又被称为第4层交换。

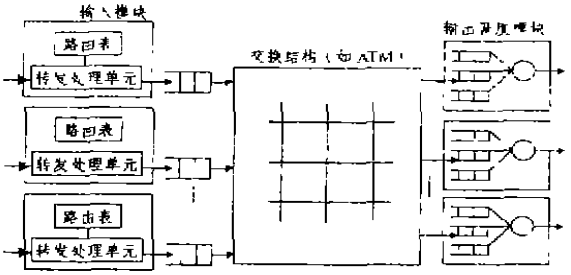

图1 交换式路由器的基本组成单元(数据通道)

图1给出了这种交换式路由器数据通道的基本组成单元。一个交换式路由器的数据通道主要由以下几个部分组成：

(1)输入模块(IP 分组头处理部分)。这部分主要进行的操作包括：IP 分组头检查和处理、路由表查询、分组识别。没有通过分组头检查的 IP 分组将被抛弃；通过查询路由表，路由器为 IP 分组找到输出端口；通过对分组进行分类，路由器把 IP 分组归并到不同的类型。

(2)交换结构。为了处理简单，绝大多数交换式路由器先将变长的 IP 分组拆分成定长的信元，然后将定长的信元通过交换结构送往目的端口，信元到达输出端口后再被拼装成变长的分组。

(3)输出调度模块，分组在到达输出端口后，通过输出调度模块送往输出链路。输出调度模块通过决定发送分组的先后顺序来保证各种服务类型的服务质量。

(4)为了防止网络发生拥塞，线路接口卡上还应该包括拥塞控制处理单元以进行拥塞控制，在必要时丢

弃分组。

上述各个部分都是数据通道上必需的部件。除此之外，还需要有一个CPU卡负责运行路由协议，生成路由表，以及管理和维护各个线路接口卡的正常工作。

4 交换式IP路由器的关键技术

4.1 路由表的快速查询算法

在因特网上，由于采用了无类别域间寻路CIDR (Classless Inter-Domain Routing)技术，网络前缀可以是变长的，路由器在分组寻路时，采用最长网络前缀匹配(Longest Prefix Match)。近年来已经提出了一些快速的查询算法，使查询速度提高到每秒几兆到几十兆分组，足以支持G比特链路了。

所有的查询算法都是在以下几个方面之间进行权衡：(1)路由表的查询速度(这主要由访问存储器的次数决定)；(2)所需要的存储器容量；(3)插入/删除表项的难易程度。

针对最长网络前缀匹配所提出的最早的算法是采用PATRICIA树，所有的基于树型结构的算法都是由这种算法变化而来的[3]。这种算法的问题对于IPv4来讲，最坏情况时需要访问32次存储器，这样的性能是无法接受的。目前的做法是采用扩展树的方法以增大存储器容量为代价来减少树的深度，从而减少存储器访问次数，加快路由表查询速度。此外，还可以采用压缩树的办法。采用压缩树的方法以后，存储每个结点的存储器容量变小。这样对存储器访问一次可以获得多个节点的信息，从而减少访问存储器的次数，加快查询速度。

从目前公开的查询算法看，查询的速度已经适应G比特链路的需要。目前的主要问题在于查询的速度越快，插入/删除表项的难度越大。

上面讨论的只是单播的情况。在对组播报文寻路时，路由器要对分组的源IP地址做最长网络前缀匹配，同时对目的组播地址做完全匹配。对这一问题虽然各厂商表示已解决，但是相应的算法没有公开。如何实现组播分组的快速转发是快速寻路算法中的一个难点。

4.2 分组识别

为了支持不同业务对服务质量的不同需求，在分组到达时路由器要对分组进行识别决定分组应接受的服务类型，用于识别分组的规则有单字段识别和多字段识别，单字段识别比较简单，常用在核心路由器上对分组进行分类，比如用IP分组头的TOS(Type Of Service)域来识别分组，而多字段识别就比较复杂，需要根据IP分组头的多个域来进行识别，例如利用五元组(源IP地址，目的IP地址，源端口号，目的端口号，传输层协议类型)来对分组进行识别，这种识别常用于边缘路由器对分组进行分类、并对IP分组的TOS域作相应的设置以备主干网的核心路由器使用。另外为了实现防火墙的功能也需要对IP分组进行识别。

4.3 交换结构

大多数交换式路由器在交换时采用定长的信元。分组在交换前被拆分成定长的信元，信元经过交换结构后在目的端口被组装成原分组，然后送往物理线路，由于这个原因，交换式路由器的交换结构同ATM交换机所采用的交换结构很相似。常用的交换结构有输入缓冲、输出缓冲和共享缓冲等方式，目前的交换式路由器基本上采用输入缓冲的方式。这主要有以下两个方面的原因。第一，采用输入缓冲方式对存储器的速度要求不高。同输出缓冲方式和共享缓冲方式相比，输入缓冲方式对存储器的要求同交换结构的规模无关，从而可以使得路由器的规模较大。第二，在因特网上TCP数据所占的比例很大。TCP数据具有突发的特点；并且TCP协议对丢包很敏感，一旦丢包发送速度就明显降低。为了保证TCP数据的业务质量，路由器内的存储容量应较大，而存储器的容量越大，其速度就越慢，这也要求采用输入缓冲方式以降低对存储器速度的要求。

4.4 调度策略

采用单个FIFO(先进先出)的输入缓冲方式存在队头阻塞。为了解决队头阻塞，大多数路由器采用了虚拟输出队列方式(VOQ：Virtual Output Queue)，即在每个输出端口为每个输出端口设置一个队列，送往同一输出端口的信元存放在同一个队列中[4]。CISCO公司的GSR12000系列路由器就采用了VOQ方式，通过仲裁算法对交换结构的连接情况进行配置。

为了支持多种服务类型，路由器对分组进行分类，将分组分成不同的类型，放入不同的队列。这样路由器在输出分组时，需要从多个队列中进行选择，决定从哪个队列中输出分组的操作称为队列调度。队列调度的主要目的是为每个数据流提供一定的服务质量保证(如带宽，时延及时延抖动)。

5 路由体系评价

从上面对路由器体系研究看出，路由器体系结构中融入了先进的分布式路由和最新的核心交换技术，其发展思想是：将路由引擎和转发引擎分开；用快速的硬件实现IP报头处理、寻径和转发；采用分布式接口；用交换结构提高各接口之间的数据通道的速度。总的来说，路由器体系结构涉及的关键技术主要是要解决与速度有关的问题、与服务质量有关的问题和与软件有关的问题。从而以这些思想和关键技术为驱动力，路

EWFQ:一种新的高速网络分组调度算法*)

EWFQ:A Novel Packet Scheduling Algorithms in High Speed Networks

任立勇　卢显良

(电子科技大学计算机学院　成都610054)

Abstract　Packet scheduling algorithm is one of crucial technologies of routers in high speed networks. In this paper we first discuss the limitation of some existing packet scheduling algorithms, then show the quantitative relationships between the GPS system and its corresponding packet WFQ system. A novel packet scheduling algorithm is proposed. It is proven to have following properties: (1) it ensures fair allocation of bandwidth among all sessions; (2) it provides deterministic delay upper bounds to a session whose traffic is constrained by a leaky bucket; (3) it has a relatively low asymptotic complexity of O(logN); (4)it has a relatively low Worst-case Fair Index(WFI). So it can be deployed in routers of high speed networks.

Keywords　Fair queueing, Scheduling algorithms, Quality of service, High speed networks

1　引言

宽带综合业务网要求能给不同的应用提供不同的服务质量(QoS),其中分组调度算法作为网络路由器中的一个重要组件起着相当关键的作用,传统的Internet是基于尽力而为(best-effort)模型实现的,该模型采取先来先服务(FCFS)的分组调度算法,这种模型具有实现简单的特点,它在假定所有应用互相协作的情况下工作得非常好,但当网络发生拥塞时,实时应用的服务质量往往得不到保证。同时,连接间的隔离性能也非常差,吞吐量大的连接得到更多的服务,某些不良行为的连接可能造成其他连接的服务质量急剧下降。

A.K.Parekh 等提出的广义处理器共享(GPS)[1]能较好地解决上述问题:1)当在数据源端实施漏桶算法的流量整形时,GPS能提供端到端的延迟界限,2)GPS能为有分组积压的连接提供公平的带宽分配。由于GPS是基于流体模型的理想化调度算法,不能用于实际系统,于是各种各样的基于分组调度的近似GPS算法相继提出[2~5],这些算法均是在连接带宽公平性分配与计算复杂度间作出平衡,如WFQ尽管提供了可与GPS相当的性能,但由于其计算复杂度较高(O(N))而不能在高速网络中运行,SCFQ算法重新定义了系统虚拟时间计算函数,将计算复杂度降低为O(1),但不足之处在于当网络连接增多时,连接的隔离和保护性能变差,同时SCFQ提供的端到端延迟较大。

为此,本文在研究WFQ与SCFQ的工作机制与不足的基础上,提出了一种新的分组调度算法EWFQ(Extended Weighted Fair Queueing),理论分析与实验证明EWFQ不仅具有计算复杂度低的特点,同时也更好地接近了GPS性能水平。

*)本文得到国家九五重点攻关项目基金资助,信息产业部生产发展基金资助。任立勇　博士研究生,主要研究方向为网络资源管理、网络协议等。卢显良　教授,博士生导师,主要研究方向为操作系统与网络应用技术。

由器的体系结构正朝着速度更快、服务质量更好和更易于综合化管理三个方向发展。

参考文献

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

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

3　Chen J S, Guerin R. Performance study of an input queueing packet switch wiich with two priority classes. IEEE Trans. Commun[J], 1991, 39(1): 117～126

4　Mckeown N. Fast Switched Backplane for a Gigabit Switched Router. White Paper, http://www.cosco.com

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

6　李津生,等编著.下一代Internet网络技术.人民邮电出版社,2001