高性能路由器中的并行包转发机制研究*)

李胜磊 张德运 刘 刚
(西安交通大学网络研究所 西安710049)

Parallel Packet Forwarding for High End Routers

LI Sheng-Lei ZHANG De-Yun LIU Gang
(Institute of Network, Xi'an Jiaotong University, Xi'an 710049)

Abstract The invention and evolution of the Dense Wavelength Division Multiplexing (DWDM) technology have brought a breakthrough to high-speed networks, and it has put a lot of pressure on research in the area of IP routers. Besides, with up-coming Quality of Service requirements raised by a wide range of communication intensive, the next-generation IP routers should be QoS-capable. In this paper, we propose a architecture called the Parallel QoS-capable IP Router (PQIR). We address one key design issue in our architecture - the distribution of IP packets to multiple independent routing agents so that the workload at routing agents is balanced and the packet ordering is preserved. We introduce the Enhanced Hash-based Distributing Scheme (EHDS)as the solution. Simulations are carried out to study the effectiveness of EHDS. The results show that EHDS does meet our design goals very well.

Keywords Router, Parallel, QoS

1 引言

Internet 的快速发展对网络带宽的要求越来越高，由于光纤技术的进步，尤其是 DWDM 技术在骨干网络上的应用，光纤能够承载的容量也越来越大。然而，根据摩尔定律，计算能力的增加是每18个月增加一倍，因此，传统的路由器就逐渐成为网络的瓶颈。而且，多媒体以及实时通信技术对服务质量也提出了严格的要求，这又增加了路由器的开销。总之，对下一代路由器的设计要求是既能够高速转发数据包，同时又能够满足服务质量的要求。

路由器的发展经历了4代。第一代的路由器采用的是中央路由单元，以及内部总线式的数据通道。线路接口卡和路由单元之间是通过内部总线连接在一起的，中央路由单元和总线是整个路由器的瓶颈。第二代路由器使用分布式路由单元取代了中央路由单元，每一个线路接口卡上都有自己的路由处理单元，总线依然是系统的瓶颈。第三代路由器采用背板总线取代了共享总线。为了满足未来的需求，第四代路由器也就是下一代的路由器将是高性能的具有 QoS 能力的路由器。通过使用新的体系结构，采用并行的办法，新的路由器的并行度和可升级能力将会更高。

人们在下一代路由器的设计上已经作了大量的研究工作。Tzi-cker 和 Prashant Pradhan[1]设计了一个称为 Suez 的基于集群技术的路由器原型。Suez 采用普通的 PC 机作为节点，通过一个10GBPS 的交换机作为系统的背板，采用集群技术协同工作。斯坦福大学的 Nick McKeown[2]的研究组合在高速交换领域也做了大量的工作。E. Basturk[3]设计了一种基于 RSVP 协议的具有 QoS 能力的交换式路由器。还有其他的许多人在高速路由、高速路由表查找、实时数据包调度、QoS 控制等领域也做了大量的研究工作[4,5]。

我们的工作同前面的工作相比，主要着重在控制调度和 QoS 问题。本论文提出了一种新的体系结构，在这种结构的基础上注重探讨了如何将数据包有效地分发到多个路由代理的调度问题。

2 高性能具有 QoS 能力的路由器的控制框架

路由器的基本功能之一是对到达的数据包做路由处理。对于一个具有 QoS 能力的路由器来说，必须做到：(1)从链路上接收数据包；(2)IP 数据包头的分析(例如数据包分类、QoS 分析)；(3)路由表查找；(4)QoS 数据包调度；(5)将数据包发送到输出链路上。为了尽可能地提高路由器的性能，这些功能应该并行地进行处理。

2.1 PQIR 体系结构

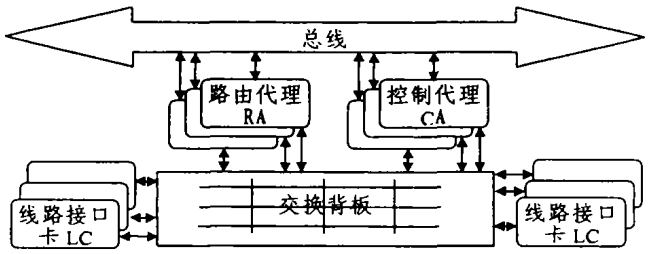

图1

PQIR 的体系结构如图1所示。它主要由下面部分组成：

·线路接口卡 LC：负责从链路上接收和发送数据包；

·路由代理 RA：负责以并行的方式执行路由表的查找；

·控制代理 CA：负责处理路由表的计算和 QoS 的控制任务；

·高速交换背板：是将 LC、RA、CA 连接在一起的高速数

*)本课题得到陕西省重大技术创新项目基金资助。李胜磊 博士研究生，主要研究领域为网络性能、路由器的研究。张德运 教授，博导，主要研究领域为网络性能、网络操作系统、网络安全。

据通道;也可以是并行化的;

·总线:广播信息的通道,例如,RA 和 CA 之间的路由更新信息。

在我们的体系结构中,我们把 IP 数据包头的分析和路由表的查找工作分开来处理,这些工作在传统的路由器设计中,通常在线路接口卡上实现。将这些工作分开来做的原因是:

·将路由表查找和 QoS 控制从线路接口卡中剥离出来可以使线路接口卡变得更加简洁高效,因此,线路接口卡可以获得更高的收发速度,满足高速链路接口的要求。

·另外,我们还可以获得更好的负载均衡的效果,提高系统的性能。因为路由查找是系统的瓶颈,我们可以通过增加路由代理的方式提高系统的路由查找能力,另外,对到达的数据包采用负载均衡的方式发送到不同的路由代理。

2.2 线路接口卡和路由代理的设计

图2表示了线路接口卡和路由代理的设计。

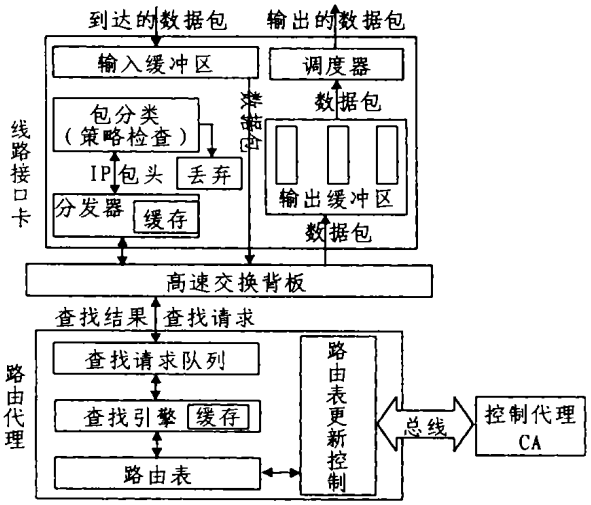

图2

当从输入链路上接收到数据包后,首先将数据包放到输入缓冲区当中。然后将 IP 数据包头剥离出来并把数据包头转发到分类器,分类器对数据包头作 DiffServ 分类或者其他的策略检查。如果数据包是非法的话,就直接丢弃。如果数据包合法的话,就由分发器根据数据包头的内容,使用特定的算法,来决定使用哪一个路由代理。决定路由代理之后,利用数据包头的信息构造路由查找请求信息,通过高速交换背板进入路由查找请求队列。路由查找引擎从请求队列中获取请求信息,并进行高速路由查找。查找完毕后,根据查找的结果修改数据包头,连同输出端口号一起,通过高速交换背板,将结果送回到线路接口卡。接收到输出端口信息后,分发器从输入缓冲区中去除原始的数据包,修改原始的数据包头,通过高速交换背板,将数据包转发到输出端口所在的线路接口卡。输出的数据包到达线路接口卡后,首先要在输出缓冲区中排队等待调度。输出缓冲区中根据优先级的不同可能有多个输出队列,我们可以使用基于 QoS 的调度算法对输出缓冲区中的数据包进行调度。

2.3 IP 数据包分发算法

我们提出了一种 IP 数据包的分发算法,由 LC 和 RA 实现,这种算法满足下面的条件:

·RA 必须工作在负载均衡的方式下,这也是我们将 RA 从 LC 中剥离出来的主要原因;

·同一个数据流的数据包必须按照到达的次序依次传送,尤其是对于 TCP 连接的数据包,这是因为乱序的 TCP 数据包可能会导致 TCP 的重传或者是超时,这样就会影响 TCP 的性能。

基于流的分发算法(Flow-based Distribution Algorithm,FDA)是一种简单的解决方案,但是这种方案具有明显的不足:可能导致负载的不均衡传输;LC 上的数据分发器必须记录每一个数据流的信息,这样在数据流很大的情况下会增加系统的开销。

Internet 的流量包括 TCP 和 UDP 流量。TCP 的流量对数据包的有序性非常敏感。我们的算法称之为基于哈希的增强型数据分发算法,可以同时满足负载均衡和数据包依序传输的要求。在我们的算法中,对于 TCP 流量来说,考虑更多的是数据流的有序性,而对于 UDP 流量来说,考虑最多的是负载均衡。

基于哈希的增强型数据分发算法

我们给每一个 RA、LC 和 CA 分配一个内部的标识号。分发 IP 数据包有一个简单的算法,称之为基于哈希的数据包分发基本算法,简称 HDA。HDA 的基本思想就是根据每一个 IP 数据包的目的 IP 地址生成一个哈希值,然后直接使用这个值作为要转发到的路由代理的 ID 号。因为具有相同目的地址的数据包被转发到相同的路由代理(相同 IP 地址的哈希值也是相同的),数据包的顺序就得到了保障。但是这种算法有两个缺点:1)尽管哈希值只是一个伪随机数字而且应该是均匀分布的,但是没有办法保证这一点,存在实在目的地址不均匀分布的情况。2)ID 号是和目的地址相关的,因此,每一个 RA 的工作负载无法动态调整,无法保证负载的均衡。

与 HDA 算法中直接使用哈希值作为路由代理的 ID 号不同,EHDA 使用的是间接哈希技术,哈希值标示的是指向哈希表的索引,哈希表中存放的是 RA 的 ID 号。哈希表中的内容(RA 的 ID 号)可以根据每一个 RA 的工作负载情况动态地调整。基本的思路就是:如果一个 RA,例如 Ra-x 正在被使用,而另一个 Ra-y 处于空闲状态,我们就相应地用 Ra-y 来替代 Ra-x,哈希表中的值也相应地被更新。图3表示了哈希表的结构以及 EHDA 算法的基本流程在 LC 端的实现。要注意的是,在这里 TCP 和 UDP 的数据流量是分开处理的,也意味着即使是相同地址的数据包也可能使用不同的 RA 进行路由查找工作。

在每一个 RA 中,工作负载被定义为路由查找请求队列的平均长度。

为了对 TCP 和 UDP 的流量进行区分处理,我们使用不同的哈希表更新策略。假设我们使用0～7表示 RA 的工作负载,一种更新概率曲线如图3所示。

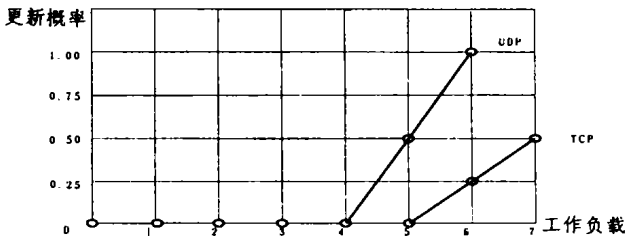

图3

在这个例子中我们可以看到,当 RA 中的工作负载达到5的时候,所有的 LC 中涉及到 UDP 部分的更新概率增加到0.5,当工作负载达到6的时候,更新概率增加到1。而当 RA 中

的工作负载达到6的时候，所有的 LC 中涉及到 TCP 部分的更新概率增加到0.25，当工作负载达到7的时候，更新概率增加到0.5。从这里我们可以看到，TCP 流要比 UDP 稳定，这样我们就在保证数据包顺序的情况下，同时获得了高负载均衡。在实际系统中，工作负载的区域和更新概率都可以根据实际情况调整。我们引进更新概率的另外一个主要原因就是为了避免工作负载的波动。例如，当 Ra-1的工作负载达到警戒线的时候，所有的 LC 都将更新哈希表，把路由查找工作切换到其它的 Ra 上，这样 RA-1的工作负载就会突然降低到0，而在另外一个时刻，所有的 LC 发现 RA-1的工作负载为0，又会全部切换回来，这样就造成了波动，是我们所不希望的。

图4表示了 EHDA 的基本工作流程。

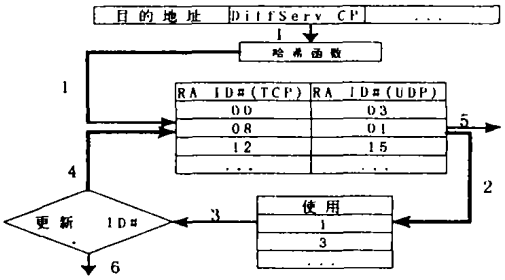

图4

(1)获取哈希值。我们使用目的地址作为键值，IP 数据包头中的其它信息，例如 DiffServ 中的 DSCP(DiffServ Code-Point)也可以作为键值使用。我们采用如下的哈希函数，假设系统中 RA 的数量为 r，并且是2的整数次幂，$r=2^q$，$q>1$，我们可以使用 CRC 码，生成多项式为 $G(x)=x^q-x-1$或者是从目的地址中随机地选择 q 位来作为哈希值。如果系统中 RA 的数量为 r，并且不是2的整数次幂，我们可以使用目的地址取模 r 的结果来作为哈希值。从理论上我们可以证明只要目的地址是均匀分布的，哈希值也是均匀分布的，这样的哈希函数可以高速地实现。

(2)获取 RA ID#。使用步骤(1)中的哈希值作为哈希表的索引，从哈希表中获取哈希值，注意在哈希表中，TCP 和 UDP 是分开处理的。

(3)检查工作负载，作出哈希表的更新决定。获取到 RA ID#后，通过查找使用表来获知当前的 RA 的工作负载情况，根据工作负载和相应的更新概率决定是否更新哈希表，注意 TCP 和 UDP 的更新概率是不同的，如果决定更新哈希表，执行步骤4，否则执行步骤6。

(4)更新哈希表。从使用表中选择工作负载最低的 RA ID#，更新哈希表。

(5)如果更新了哈希表，就使用新的 RA ID#作为最后的结果。

(6)如果未更新哈希表，就使用原先的 RA ID#作为最后的结果。

在实际实现中，由于哈希表和使用表都很小，我们可以把它们放到 LC 的缓存中，这样可以提高系统的性能。

3 仿真和结果

为了检验我们的设计方案，我们对系统进行了仿真测试，并对测试结果进行分析。我们设计了一个基本的事件驱动的仿真器，仿真的结果证明设计是可行的。

A. 仿真模型

我们使用 TCP 和 UDP 的数据流作为输入流量，每一个数据流都有自己的起始时间 t_{start}，结束时间 t_{end}和发送速率 r，数据流的输入符合泊松分布。发送速率在一定的范围内随机选择。在我们的实验中，我们固定地使用4个 LC，8个 RA，使用 HAD 和 EHDA 最对比试验。

B. 仿真结果和实验分析

我们测试的重点是 EHDA 算法的有效性。我们使用图4表示的曲线作为哈希表的更新策略，试验结果如表1所示：

表1

输入流量类型	HDA 算法丢包率	EHDA 算法丢包率	EHDA 哈希表更新次数 TCP	EHDA 哈希表更新次数 UDP
均衡输入	3.91%	0.1%	1079	5561
非均衡输入	54.75%	0.15%	1178	5460

从上表我们可以看出，无论是均衡还是非均衡输入流量，EHDA 算法的效果都要比 HDA 算法好，而且哈希表的更新和输入流量模式无关(均衡或非均衡)，它只依赖于更新策略，我们可以通过调整更新策略来获取不同程度的对 TCP 流量有序性的保证，这样 EHDA 算法也可以保证 TCP 的有序性。

在大输入流量的情况下 RA 的计算能力不能满足输入流量的要求，丢包率很高，在系统中增加了7个 RA 后，升级后的系统可以处理输入的流量要求，丢包率降为0，这就证明了系统的可扩充性良好。

从上面的结果我们可以得出这样的结论：无论是怎样的输入流量，RA 可以做到均衡工作，TCP 的包顺序可以得到保证，保证程度可以通过更新策略进行调整。系统的可扩充性可以得到保证，可以通过增加系统的 RA 来增加系统的处理能力。

结论 本论文提出了一种新的 IP 路由器的体系结构，在此体系结构的基础上，设计了一种新的可以并行工作的 IP 数据包分发算法，并对系统进行了仿真研究，结果证明我们的设计是可行的。

参 考 文 献

1 Chiueh Tzi-cker, Pradhan P. Suez: A Cluster-Based Scalable Real-Time Packet Router. In: Proc. of 20th Intl. Conf. Distributed Computing Systems, Taipei, Taiwan, April 2000. 136～144

2 McKewn N. A Fast Switched Backplane for a Gigabit Switched Router. Business Communications Review, http: // www. bcr. com/bcrmag/12/mckeown. htm, March/April 1997

3 Basturk E, et al. Design and Implementation of a QoS Capable Switch-Router. Computer Networks, 1999, 31(1-2): 17～30

4 Wang Jun, Nahrstedt K. Parallel IP Packet Forwarding for Tomorrow's IP Routers. In: Proc. of 2001 IEEE Workshop on High Performance Switching and Routing (HPSR'01), Dallas, TX, May 29-31, 2001. 253～257

5 Gupta P, Lin S, McKeown N. Routing Lookups in Hardware at Memory Access Speeds. In IEEE Infocom 1998, San Francisco, volume 3, April 1998. 1240～1247