

一个改进的路由器包转发模型

李小勇, 刘东喜, 陈 凯, 梁阿磊, 白英彩

(上海交通大学计算机科学与工程系, 上海 200030)

摘 要: 本文提出并实现了一个改进的路由器包转发模型, 该模型解决了传统模型在系统过载时吞吐量下降和延迟剧增的问题, 并避免了传统模型中多级队列导致的包转发延迟. 实际测试表明路由器的性能得到显著提高, 证明了改进模型的优越性.

关键词: 路由器; 任务调度; 吞吐量; 延迟; 排队模型

中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2002) 12A-2158-03

An Improved Router Packet Forwarding Model

LI Xiao-yong, LIU Dong-xi, CHEN Kai, LIANG A-lei, BAI Ying-cai

(Department of Computer Science and Engineering, Shanghai JiaoTong University, Shanghai 200030)

Abstract: This paper presents an improved router packet forwarding model. It solves the problems of throughput decreasing and latency increasing in the traditional packet forwarding model, and avoids the latency introduced by multi-level queues in the traditional model. The practical test shows that the performance of the router is improved remarkably and proves the advantages of the improved model.

Key words: router; task scheduling; throughput; latency; queuing model

1 引言

基于主机或软件的路由器, 其物理接口支持广泛, 应用程序丰富, 易于开发和扩展新的功能, 所以经常用作网络接入设备和新技术实验平台, 在实践中具有重要的地位和作用. 如何设计高效的软件系统来提高路由器的性能, 始终是近年来的研究热点之一. 目前已提出多种提高路由器性能的措施, 如设计高效的路由表组织和查找算法^[4,10]; 提高 CPU Cache 的利用率^[8]; 以及利用“network flow”现象设计路由表缓冲和快速转发^[6,7,9], 对路由器软件结构进行优化^[1,2]等.

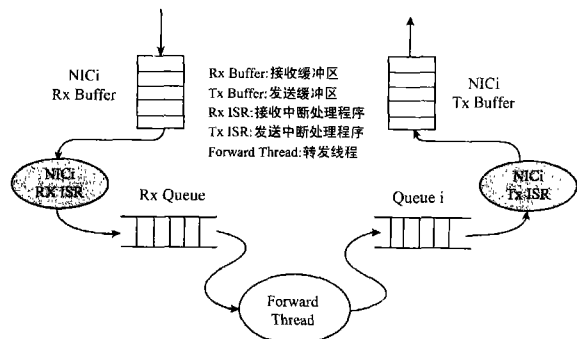


图1 传统路由器包转发模型

虽然上述各种优化措施能够有效地提高路由器的吞吐量, 但无法解决系统过载时吞吐量下降和延迟剧增的问题. 本

文提出一个改进的路由器包转发模型, 该模型通过引入消息机制, 改进了系统的任务调度, 简化了包转发过程中的通信队列, 保证 CPU 资源在包转发过程中的多个任务之间能够合理分配, 从而消除了传统模型在系统过载时的性能下降问题, 并有效地提高了路由器的最大吞吐量, 减小了包转发延迟.

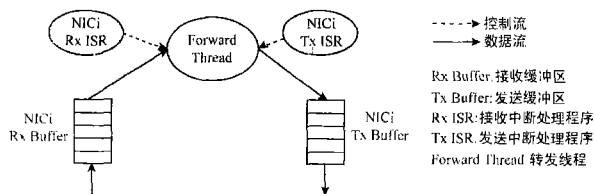


图2 改进路由器包转发模型

2 传统路由器包转发模型

基于软件的路由器中有多个网络接口卡 NIC, 完成从传输介质接收和发送包、解码、编码的工作, 它们通过共享总线的方式和中央处理器 CPU 进行通信. CPU 完成 NIC 的驱动、路由表的维护和包的转发工作.

图 1 为传统路由器包转发模型, Linux、NetBSD 等网络操作系统的包转发机制都对应于此模型. 在传统模型中, 包转发要经过 Rx ISR、Forward Thread 和 Tx ISR 三部分处理, 其中 Rx ISR、Tx ISR 以中断方式运行, Forward Thread 以内核线程方式运行, 它们的优先级关系为: $P_{Rx\ ISR} > P_{Tx\ ISR} > P_{Forward\ Thread}$.

Rx ISR 运行的时间和接口的包到达率有关,随之增加而增加. 由于 Rx ISR 具有最高优先级,当包到达率很高时, Rx ISR 将占用大量 CPU 资源,使得 Forward Thread 和 Tx ISR 的运行时间减少,接收队列 Rx Queue 溢出,系统处于“接收-丢弃”的恶性循环中,其必然导致路由器吞吐量下降,包转发延迟增加.

3 改进模型及性能分析

为了消除传统模型的缺陷,本文提出一个改进的路由器包转发模型,如图 2. 在改进模型中, Rx ISR、Tx ISR 只发送控制消息,唤醒 Forward Thread; 取消了接收队列 Rx Queue, Forward Thread 调用接口驱动程序直接从每个网络接口对应的接收缓冲区读取包,依次完成包的接收、转发、发送缓冲区的释放等全部工作. Rx ISR、Tx ISR 发送控制消息后,关闭中断,当 Forward Thread 被唤醒,完成处理工作后,打开相应中断.

改进模型通过将包转发过程中的多个处理任务合并到一个线程中,避免了单个任务独占 CPU 资源,从而消除了系统过载时的吞吐量下降问题;减少了包转发过程中的队列数,避免了传统模型采用多级队列所导致的延迟问题;改进模型也减少了系统任务调度和中断的开销,必然有助于提高路由器的最大吞吐量.

下面对改进模型和传统模型的吞吐量以及延迟特性进行对比分析.

3.1 吞吐量分析

在忽略系统调度和中断开销的情况下, Rx ISR、Forward Thread、Tx ISR 处理的包数满足如下约束关系:

$$\sum_{i=1}^n n_{ri} * C_r + n_f * C_f + \sum_{i=1}^n n_{ti} * C_t \leq C \quad (1)$$

由于 Rx ISR、Forward Thread、Tx ISR 之间依次为生产者-消费者关系, Tx ISR 的优先级高于 Forward Thread 的优先级,它们处理的包数满足如下关系:

$$\sum_{i=1}^n n_{ri} \geq n_f = \sum_{i=1}^n n_{ti} \quad (2)$$

显然当式(2)中等号成立时,路由器的吞吐量最大,即:

$$\sum_{i=1}^n n_{ri} = n_f = \sum_{i=1}^n n_{ti} \quad (3)$$

且最大吞吐量为:

$$S_{\max} = \frac{C}{C_r + C_t + C_f} \quad (4)$$

在以上各式中,各参量的意义为:

n : 路由器接口数;

C_{ri} : NIC_i Rx ISR 处理单个包需要的 CPU 操作数;

C_f : Forward Thread 处理单个包需要的 CPU 操作数;

C_{ti} : NIC_i Tx ISR 处理单个包需要的 CPU 操作数;

C : CPU 处理能力(即 CPU 单位时间内可执行的操作数);

n_{ri}, n_{ti} : NIC_i Rx ISR、Tx ISR 分别在单位时间内处理的包数;

n_f : Forward Thread 在单位时间内处理的包数.

在传统路由器包转发模型中, $\sum_{i=1}^n n_{ri}$ 受到各个接口的包

到达率的影响,随包到达率的增加而增加. 由式(1)可知,在传统模型中,当 $\sum_{i=1}^n n_{ri} \geq S_{\max}$ 时,随着 $\sum_{i=1}^n n_{ri}$ 继续增加, $\sum_{i=1}^n n_{ti}$ 必然减小,在路由器性能上表现为:当包到达率超过路由器的最大吞吐量时,随着包到达率继续增加,路由器的吞吐量下降. 在改进模型中,通过将包处理任务合并为一个内核线程,消除了包转发过程中包到达率对 CPU 资源调度的影响,保证了优化条件式(3)成立,使得路由器始终工作于最优状态,保证在系统过载时,吞吐量不会随着包到达率的增加而下降.

3.2 延迟特性分析

传统的路由器包转发模型和改进的路由器包转发模型所对应的排队模型分别如图 3(a)、(b)所示. 为简化分析,假定:

(1) 路由器接口类型相同,从而有 $C_{r1} = C_{r2} = \dots = C_{rn} = C_r$, $C_{t1} = C_{t2} = \dots = C_{tn} = C_t$;

(2) 每个接口的包到达率呈泊松分布,且 $\lambda_1 = \lambda_2 = \dots = \lambda_n = \lambda$;

(3) 每个处理任务和对应的输入队列为一个 $M/M/1$ 系统.

在图 3(a)所示传统路由器包转发排队模型中,有:

$$\mu_{r1} = \mu_{r2} = \dots = \mu_{rn} = \mu_r = \frac{C}{n} * \frac{1}{C_r + C_f + C_t} \quad (5)$$

$$\mu_{f1} = C * \frac{1}{C_r + C_f + C_t} \quad (6)$$

$$\mu_{t1} = \mu_{t2} = \dots = \mu_{tn} = \mu_t = \frac{C}{n} * \frac{1}{C_r + C_f + C_t} \quad (7)$$

则路由器在丢包率为零时的报文转发延迟 T_{tran} 为^[3]:

$$T_{\text{tran}} = \frac{1}{\mu_r - \lambda} + \frac{1}{\mu_{f1} - n * \lambda} + \frac{1}{\mu_t - \lambda} = \frac{(2n+1)(C_r + C_f + C_t)}{C - n\lambda(C_r + C_f + C_t)} \quad (8)$$

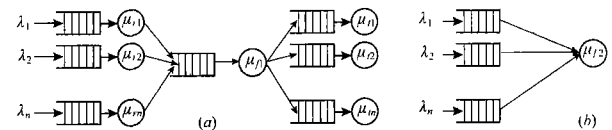


图 3 路由器包转发排队模型 (a) 传统路由器包转发排队模型; (b) 改进路由器包转发排队模型

在图 3(b)所示改进路由器包转发排队模型中,

$$\mu_{f2} = \frac{C}{C_r + C_f + C_t}$$

丢包率为零时的报文转发延迟 T_{imp} 为:

$$T_{\text{imp}} = \frac{1}{\mu_{f2} - n\lambda} = \frac{C_r + C_f + C_t}{C - n\lambda(C_r + C_f + C_t)} \quad (9)$$

由式(8)、(9)可知,在总的包到达率 $n * \lambda$ 为常量时,传统模型包转发延迟和接口数 n 呈正比. 在改进后的模型中,延迟降低为传统模型的 $1/(2n+1)$.

4 实验测试结果

为了检验改进模型对路由器性能产生的影响,对早期设计的一个路由器原型系统 BQR 进行了吞吐量和延迟特性测试^[5]. BQR 的 CPU 为多功能通信处理器 MPC860, 时钟频率

25MHz,操作系统为实时操作系统 VRTX.测试设备选用网络设备测试仪 SmartBits 2000,包流量模型设置为确定性流量,包长度为 64 字节.测试结果见图 4.(在图中,Before Optimization 对应于传统路由器包转发模型,After Optimization 对应于改进的路由器包转发模型,pps 为 packets per second).

由图 4(a)可看出,在传统模型中,路由器的最大吞吐量 MLFRR(Maximum Lose Free Receive Rate)^[5]为 2.6Kpps,当包到达率超过 MLFRR 后,吞吐量就会急剧下降.采用改进模型后,MLFRR 提高到 3.9Kpps,即使包到达率继续增大,吞吐量仍然能够保持在此水平,没有随之而下降.

图 4(b)表明,在传统模型中,包转发延迟随着包输入速率的增大而持续增加,在输入速率接近 MLFRR 时,延迟急剧增加.采用改进模型后,包到达率小于 MLFRR 时,包转发延迟约为 270 μ s,在包到达率大于 MLFRR 时,延迟约为 4.1ms.和优化前相比,包转发延迟显著降低,而且,延迟抖动也显著减小,这对网络中

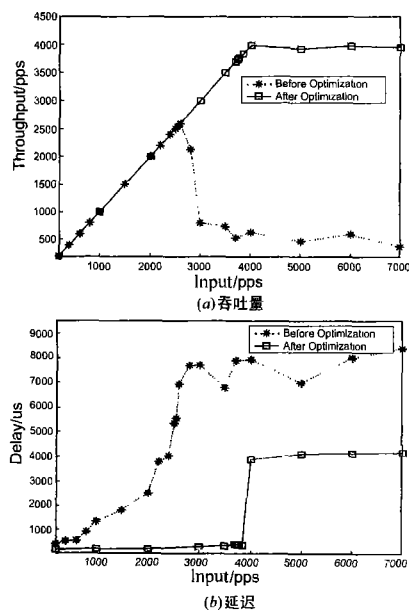


图 4 优化前后路由器吞吐量、延迟特性比较

的视频和音频业务的支持是十分重要的.

在图 4(b)中,我们也可以看出,即使采用改进模型,当包输入速率大于 MLFRR 时,延迟也会急剧增加.这是因为在输入速率小于 MLFRR 时,包转发平均延迟为单个包的处理时间,而当输入速率大于 MLFRR 后,包延迟将增加 k 倍, k 为网络接口对应的接收缓冲区的数目.显然,当增加接收缓冲区数目时,路由器对突发流量的缓冲能力得到改善,但是当包到达率超过 MLFRR 时,延迟也将随之增加.

5 结束语

本文针对传统路由器包转发模型的不足,提出一个改进的路由器包转发模型.改进模型通过引入消息机制,简化包转发过程中的通信队列,改进了系统的任务调度,消除了包转发过程中包到达率对 CPU 资源调度的影响,使 CPU 资源能够在包转发过程中的多个任务间合理分配,解决了传统模型在系

统过载时的性能下降问题,有效降低了包转发延迟.由于改进模型减少了系统任务调度和中断的开销,使路由器的最大吞吐量也得以显著提高.

本文的分析和实验说明,路由器的性能不仅和各个模块局部的效率有关,也与系统的任务调度机制有密切关系,通过优化任务调度机制和通信队列,可以有效地提高路由器的性能.

参考文献:

- [1] D Decasper, Z Dittia, V Parulkar, B Plattner. Router plugins: a software architecture for next-generation routers[J]. IEEE/ACM Transactions on Networking, 2000, 8(1): 2-15.
- [2] Eddie Kohler, Robert Morris, Benjie Chen, John Jannotti, M Frans Kaashoek. The click modular router[J]. ACM Transactions on Computer Systems, 2000, 18(3): 263-297.
- [3] Leonard Kleinrock. Queueing Systems. Volume I: Theory[M]. Chichester: John Wiley & Sons Inc. 1975.
- [4] M Degermark, A Brodnik, S Carlsson, S Pink. Small forwarding tables for fast routing lookups[J]. Computer Communication Review, 1997, 27(4): 3-14.
- [5] S Bradner, J McQuaid. rfc2544: Benchmarking Methodology for Network Interconnect Devices[S]. 1999.
- [6] S Walton, A Hutton, J Touch. High-speed data paths in host-based routers[J]. IEEE Computer, 1998, 31(11): 46-52.
- [7] Tung Ye, Che Hao. A flow caching mechanism for fast packet forwarding[J]. Computer Communications, 2002, 25(14): 1257-1262.
- [8] Tzi-cker Chiueh, Prashant Pradhan. High performance IP routing table lookup using CPU caching[A]. Proc. of IEEE INFOCOMM[C]. California: IEEE Computer Society Press, 1999. 1421-1428.
- [9] Vijay Bollapragada. Cisco IOS 精髓[M]. 邓劲生,等译.北京:中国电力出版社, 2001. 59-66.
- [10] V Srinivasan, G Varghese. Fast address lookups using controlled prefix expansion[J]. ACM Transactions on Computer Systems, 1999, 17(1): 1-40.

作者简介:

李小勇 男, 1972 年生于甘肃天水, 1999 年毕业于电子科技大学, 获信息与电子系统硕士学位, 现为上海交通大学计算机系博士研究生, 主要研究方向为高速网络、网络安全. Email: xyli_sohu@sohu.com.

刘东喜 男, 1973 年生于山西稷山, 1999 年毕业于太原理工大学, 获计算机应用硕士学位, 现为上海交通大学计算机系博士研究生, 主要研究方向为形式化验证技术、安全协议.

陈凯 男, 1973 年 11 月生于江西南昌, 1998 年毕业于华中理工大学通信与电子系统专业, 获硕士学位, 现为上海交通大学计算机系博士研究生, 主要研究方向为路由器体系结构、网络存储、网络安全.