ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.3, March 2008, pp.733−743
DOI: 10.3724/SP.J.1001.2008.00733

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

路由器缓存需求*

李玉峰[1,2+], 邱 菡[1], 兰巨龙[1], 汪斌强[1]

[1](国家数字交换系统工程技术研究中心,河南 郑州 450002)

[2](防空兵指挥学院 信息控制系,河南 郑州 450052)

Buffer Sizing in Internet Routers

LI Yu-Feng[1,2+], QIU Han[1], LAN Ju-Long[1], WANG Bin-Qiang[1]

[1](National Digital Switching System Engineering & Technological Research Center, Zhengzhou 450002, China)

[2](Department of Information and Control, Air Defense Command College, Zhengzhou 450052, China)

+ Corresponding author: Phn: +86-371-63532852, Fax: +86-371-63941700, E-mail: lyf@mail.ndsc.com.cn

Li YF, Qiu H, Lan JL, Wang BQ. Buffer sizing in Internet routers. *Journal of Software*, 2008,19(3):733−743.
http://www.jos.org.cn/1000-9825/19/733.htm

Abstract: A brief summary of the effect of router buffers on network performance is presented. The previous buffer sizing works based on queuing theory are analyzed under two classes of traffic models, the models exhibiting long-range dependence and the models exhibiting short-range dependence. Another type of buffer sizing works are based on TCP models, and the rule of thumb, the small buffers rule, the drop-based buffers rule and the tiny buffers rule are the four main recent results in these works. The results of the four rules and some preliminary experiments to validate these rules are carefully summarized. Research directions and open problems are also discussed.
Key words: router; buffer size; TCP; queuing theory

摘 要: 综述了路由器缓存的作用.基于随机服务理论,分析了长相关流量模型和短相关流量模型输入路由器系统时的各种缓存分析结论,着重论述了基于 TCP 协议模型的各种最新缓存需求研究成果,包括经验法则、小缓存法则、基于丢包率的缓存法则和极小缓存法则.总结了以往缓存研究的不足,并提出了下一步的研究方向.
关键词: 路由器;缓存需求;TCP;排队论
中图法分类号: TP393 **文献标识码**: A

作为不同网络之间互相连接的枢纽,路由器系统构成了基于 TCP/IP 的国际互联网络的主体脉络,是 Internet 的骨架.如图 1 所示,因特网流量平均以每半年翻一番的速率快速增长,以超摩尔定律发展的光传输能力,正在沿着每 7 个月传输带宽增大 2 倍的速率发展,而骨干网路由器容量的发展由于受摩尔定理所限,其增长速度仅为每 18 个月增大 2.2 倍,从而使得网络发展的瓶颈逐渐集中到路由器节点上.也可以说,路由器的性能直接影响网络的性能,是网络通信的最主要瓶颈之一.对于高速路由器的性能评估方法,RFC1242[1]和 RFC2544[2]进行了详细说明,其主要系统性能指标有吞吐率(throughput,整体输入流速率/输出速率)、利用率(utilization,平

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2005AA121210 (国家高技术研究发展计划(863)); the National Basic Research Program of China under Grant No.2007CB307102 (国家重点基础研究发展计划(973))
Received 2007-06-19; Accepted 2007-09-04

均输入速率/最大输出速率)、丢包率(loss rate)、延迟(delay)、缓存大小(buffer size)和实现复杂性(complexity of implementation)等.其中,缓存大小还直接影响其他性能指标,具有特殊的地位.若路由器设置大容量的缓存,有利的方面是能够更好地吸收链路上突发性变速率到达业务,当链路上发生拥塞时能够对新进入的数据包进行缓存,从而降低丢包率,维持高链路利用率;不利的方面是过大的缓存将导致延迟的增大,增加实现的复杂度.与此相对的是,小的缓存能够降低延迟和实现复杂度,但是利用率、丢包率和吞吐率指标将有所损失.因此,研究路由器的缓存需求是路由器优化设计的基础,是提高路由器整体性能进而提高整个网络性能的关键,然而,到目前为止,关于路由器的缓存容量需求问题仍未有统一的结论[3,4].

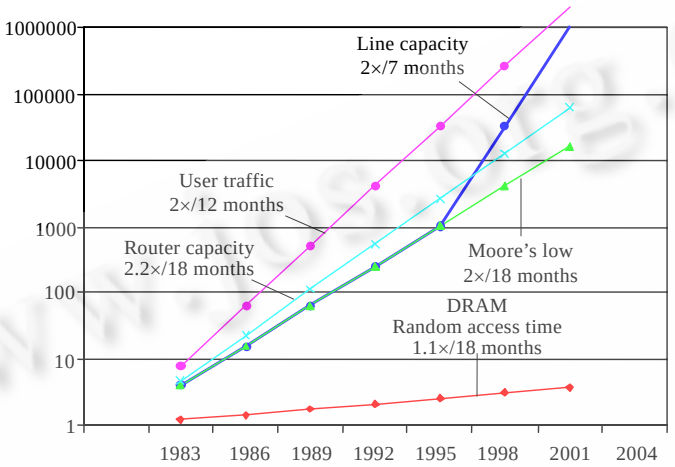

Fig.1 Trends in technology, routers and traffic

图 1 器件技术、路由器和网络流量的发展趋势图

已有学者对路由器的缓存需求进行了相关研究,概括地说,这些研究包括两大类分析方法:

第一,利用随机服务理论,也称为排队论的方法.即给定输入流量模型、服务模型,利用排队理论分析排队的队长分布(缓存大小),计算丢包率、时延等关键性能参数.这类方法不考虑输入流量的网络反馈(网络反馈包括由所观察到的链路往返时间和丢包率来触发的 TCP 自适应窗口机制和路由器中的拥塞避免策略),是一种开环方式的分析方法.该方法又可以粗略地分为两个方向:一是基于 Markov 的输入过程来得到关于缓存大小和丢包率、平均等待时间等性能参数的关系,由于输入流量模型的共同特点是只能描述网络流量序列的短时相关性,因此,本文称其为短相关(short-range dependence,简称 SRD)排队分析法;另一个是研究自相似/长相关流量输入时系统的渐进排队性能,称为长相关(long-range dependence,简称 LRD)排队分析法.

第二,以传输控制协议——TCP 协议模型为基础进行分析.当前骨干网络中,基于 TCP 的流量占到网络总流量的 85%~95%,UDP 所占比例不足 10%,其他类型协议所占比例更少[5],鉴于 TCP 流量在网络中的主导地位,众多学者基于 TCP 协议模型探讨了路由器的缓存容量需求,比较有代表性的分析结论包括:经验法则(rule of thumb)、小缓存法则(small buffer rule)、极小缓存法则(tiny buffer rule)和基于丢包率的缓存法则(drop-based buffer rule).由于 TCP 机制本身是一种闭环的反馈系统,因此,该类分析方法本质上是一种闭环分析方法.

本文后续内容将首先对两大类缓存分析方法的研究成果进行分析、总结和比较,随后对路由器缓存研究的其他相关成果进行介绍,最后对全文进行总结,并提出进一步的研究方向.

1 基于随机服务理论的缓存分析-开环分析法

目前,排队论方法被认为是研究网络性能的主要分析方法,大量基于排队论的研究推动了高速路由器的设计与发展.排队系统包括 3 个基本组成部分:输入过程、排队规则和服务.在研究网络性能时,输入流量模型是最为重要的组成部分.

1.1 SRD特性流量输入时的路由器缓存需求分析

众所周知,话务模型在电信网络规划与设计中发挥了重要的作用,当时,为了计算程控交换机的容量与阻塞概率等关系,提出了实用的话务模型,即爱尔兰(Erlang)模型.受此启发,随着ARPAnet的发展,自20世纪六七十年代开始,以Kleinlock为代表的研究人员开展了计算机网络流量模型及性能评价等方面的研究工作,取得了一系列的研究成果.由于受到落后的网络测量技术限制,大部分性能评价工作对流量的统计特性采用人为假设,在很长一段时间内都采用传统的话务模型或其改进形式来研究网络流量特性.ATM 技术的发展促进了网络流量建模及性能研究的进一步发展,最初也采用了 Poisson 过程对网络流量特性进行拟合.随着研究的深入,逐步引入了一些较为复杂的随机模型,如 fluid-flow 模型、packet-train 模型、马尔可夫调制的 Poisson 过程(Markov-modulated Poisson process,简称 MMPP)以及批到达 Markov 过程等.这些模型的共同特点是只能描述网络流量序列的短时相关性 SRD,即网络流量数据的自相关函数随着警方间间隔的增大呈指数趋势衰减.当时间尺度增加时,在统计意义上,单位时间内将趋于白噪声.

在用 SRD 特性的业务模型分析 IP 网络排队系统的性能时,对于一个服务速率为 C_L packet/s 的排队系统,缓冲器容量为 L,队列长度为 n,则排队系统溢出概率 $\Pr[n>L]\sim e^{-\delta L}$(其中,$\delta$为大于0的系数,与输入业务参数和 C,L 有关),即溢出概率与缓冲器容量呈负指数关系.此时,丢包率随缓存的增加快速降低并趋近于0,小数量的缓存便可以满足丢包率的要求,而且小缓存还能降低系统的排队时延,降低时延抖动.因此,在到达报文符合 SRD 的条件下,小缓存便可以满足路由器性能指标要求.

1.2 LRD特性流量输入时的路由器缓存需求分析

人们经过测量及分析[6,7]发现,许多分组交换网络(包括局域网、广域网、公共信道信令网、ATM 网)中的信息流具有长相关性 LRD/自相似性(self-similar).由于自相似特性比传统 Poisson 特性更难于进行排队理论研究,而且作为突发流量的一种建模方法,对流量突发特征的参数描述也远未能达成一致,因此,通过快速生成具有自相似特性的业务源进行排队仿真分析研究是目前一种有效的研究方法.现有的自相似业务生成方法主要有:基于分形高斯噪音(fractional Gaussian noise,简称 FGN)和分形布朗运动(fractional Brownian motion,简称 FBM)的业务生成方法;基于 *FARIMA*(*p*,*d*,*q*)过程的业务生成方法、基于混沌映射的业务生成方法、直接叠加具有重尾特性的 ON/OFF 业务源等.

在排队系统的快速仿真中,重点采样(important sampling,简称 IS)方法已成为一种重要的工具[8].Huang 等人对自相似业务下的排队性能进行了 IS 仿真,结果表明,传统流量模型对排队性能的估计过于乐观[9].Erramilli 等人利用以太网采样数据驱动的仿真实验对流量建模进行了研究,Bellcore 的 Neidhardt 和 Wang 分析了多时间尺度自相似特性对排队分析的作用[10],Georgians 等人对自相似过程输入时,排队系统的缓冲区溢出进行了研究[11].他们的研究结果都表明,随着缓冲区的增加,溢出概率并非按负指数方式迅速下降,而是下降得很慢.

Norros[12,13]根据分形布朗运动 FBM 提出了正则自相似模型,他利用大偏差技术推导出稳态下队列溢出概率服从韦布尔(Weibull)分布,Duffield 等人也给出了相同的结论[14].在此基础上,Rananand 给出了长相关过程输入时,队列溢出概率的一个更紧的上界[15],并进行了数值计算.

Erramilli 等人[16]利用混沌映射对自相似业务进行了研究.该方法的基本思路是把状态空间扩展为一个闭集,通过一个非线性的混沌映射 $f(\cdot)$来描述连续状态变量在离散时间上的变化:当状态变量超过某门限值时,业务源以峰值速率产生分组(对应于 ON 状态);而当状态变量小于该门限值时,业务源就不会产生分组(对应于 OFF 状态).通过选择适当的映射 $f(\cdot)$,该方法会产生一系列具有不同特性的 ON/OFF 业务源模型,使得 ON 与 OFF 状态的停留时间分别服从不同特性的分布.这种构造方法可以使研究人员通过解析复合系统的非线性方程来获得有关排队系统的瞬态和稳态性能,研究人员可以通过选取不同 ON/OFF 特性的业务源,对相应的排队性能进行评估.Pruthi[17]利用此方法得到了以下结论:(1) ON 周期具有轻尾分布、OFF 周期具有重尾分布的单个业务源,使得队列溢出概率呈指数衰减;(2) ON 和 OFF 时长都服从重尾分布的单个业务源使得队列溢出概率具有幂函数特性.

1.3 小 结

综上,在SRD模型输入排队系统时,溢出概率与缓冲器容量呈负指数关系,小缓存即可满足路由器丢包率需求,同时还具有较好的时延特性.LRD特性网络流量输入时,排队模型产生Weibull或多项式分布的排队性能,路由器需大容量的缓存才能满足丢包率特性,这与基于SRD排队分析的结果大不相同.为说明问题,引入文献[18]中利用蒙特卡罗仿真方法给出的典型对比结果,如图2所示.

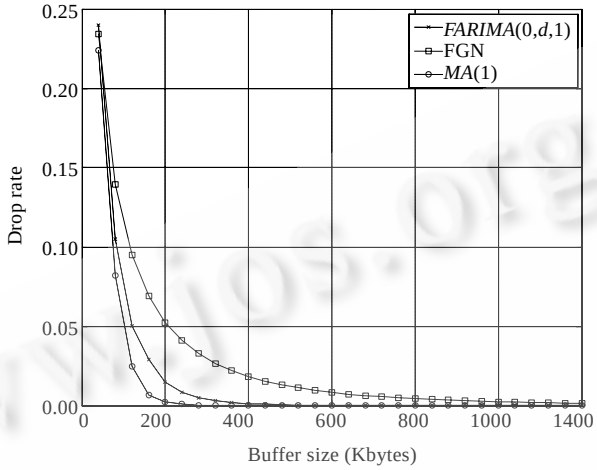

Fig.2 Drop rate vs. buffer size under different traffic models

图2 不同流量模型下丢包率与缓存的关系

从图2中可以看到,在满足相同丢包率要求的条件下,长相关和短相关性质会对路由器的缓存容量需求产生不同的影响.

当缓存小于100Kbytes时,随着缓存的增加,各种输入情况下的包丢失率都呈陡峭的直线减小,*FARIMA*(0,*d*,1)和*MA*(1)输入下的丢包率非常接近,而FGN输入的丢包率明显大于前两种情况.这说明,在缓存较小时,长相关(FGN)的丢包率大于短相关*MA*(1)的丢包率,而同时存在两种性质的模型*FARIMA*(0,*d*,1)输入时,则其短相关性主要决定丢包率,其长相关性对丢包率的影响非常小.

当缓存超过100Kbytes以后,随着缓存的增大,短相关*MA*(1)下丢包率的减小比*FARIMA*(0,*d*,1)下的减小速率要大,而FGN的减小比前两种情况都较缓慢.这说明随着缓存的增大,长相关性逐渐对丢包率产生影响,从而使*FARIMA*(0,*d*,1)和*MA*(1)下的情况有了区别.

当缓存超过250Kbytes以后,随着缓存的增大,*MA*(1)下的丢包率几乎可以接近0,而*FARIMA*(0,*d*,1)下的丢包率则逐渐减小,到600Kbytes以后,与*MA*(1)下的丢包率相近;而FGN的减小比前两种情况都缓慢,并且可以看出近似为Weibull分布.这说明随着缓存的进一步增大,FGN的长相关性使丢包率呈缓慢减小的趋势.

2 基于TCP协议模型的缓存分析-闭环分析法

本文将基于TCP协议模型的缓存需求研究成果归结为以下几类:经验法则,小缓存法则,基于丢包率的缓存法则,极小缓存法则.我们首先介绍各种法则,然后进行小结.

2.1 经验法则

- 分析假设和结论

1994年,Villamizar和Song在文献[19]中提出了著名的路由器缓存设置法则——经验法则(rule-of- thumb),该法则假设拥塞链路上存在一条长TCP流(从未离开过慢启动阶段的TCP流为短TCP流,反之,则为长TCP流),由于TCP窗口在稳态下呈锯齿形变化,要维持拥塞链路100%的利用率,拥塞缓存大小应与带宽时延积相等

(bandwidth delay product,简称 BDP),即 $B=RTT\times C$,其中,B 为拥塞路由器所需的缓存,RTT(round trip time)为一个 TCP 连接的平均往返时间,C 为拥塞链路的带宽.

- 经验法则的验证和讨论

文献[19]中分别对 1 条、4 条和 8 条 TCP 流通过容量为 40Mb/s 的链路进行了测试,通过改变缓存的大小,观察被测链路的利用率,对经验法则进行了验证.结果表明:经验法则不仅在单条 TCP 长流条件下成立,而且在流数目为 4 和 8 时依然成立.成立的原因在于:在少量 TCP 流数目条件下,由于各个流的丢包几乎同时发生,从而使各个 TCP 窗口的锯齿变化趋于同步,导致了聚合窗口依旧遵循锯齿变化,缓存需求不变.

经验法则多年以来一直指导着路由器的设计.按照这一法则,路由器可获得 100%的链路利用率,从而可以有效地利用昂贵的传输链路.但随着核心路由器的线路接口速率的快速提高,继续遵循这一法则会给路由器的设计带来相当大的困难.目前,商用路由器中,TCP 连接的 RTT 约为 250ms[20],这就意味着当链路速率为 40Gb/s 时,需要 10Gbits(1.25Gbytes)的缓存规模.实现如此大的缓存,若使用功耗为 250mW/Mbit、容量为 32Mbits 的静态随机存取存储器(static random access memory,简称 SRAM),那么,40Gb/s 的线卡对 SRAM 需求在 300 片以上,总功耗高达 2.5kW;若换用功耗为 4mW/Mbit、容量为 1Gbit 的动态随机存取存储器(dynamic random access memory,简称 DRAM),10 片就能满足要求,其功耗仅为 40W.但这不仅增加了路由器设计的复杂度、成本和功耗,而且当拥塞发生时,还将增大端到端的延迟.

2.2 小缓存法则

- 分析假设和结论

当今骨干网络的链路与 1994 年相比已有很大的不同,大量 TCP 流(flow)同时共享一条骨干链路,Fraleigh 在文献[21]中指出,在骨干网络,一条 2.5Gb/s(OC48c)或 10Gb/s(OC192)的链路在同一时刻承载的流数量通常会多于 10 000 条.鉴于此,对缓存的分析应基于大规模的 TCP 流进行.

2004 年,Appenzeller 在文献[22]中提出了较小缓存法则(small buffer rule),认为骨干链路要获得 100%的链路利用率,路由器所需缓存仅需满足

$$B=\frac{RTT\times C}{\sqrt{N}},$$

其中,N 代表拥塞链路上共享的长 TCP 流数目.

小缓存法则认为:当足够大的长 TCP 流数目共享一条链路时,这些长 TCP 流的传输将非同步化,根据中心极限定理,汇聚后的拥塞窗口的变化服从正态分布;当 $N\geq 500$ 时,同步发生的概率低于 10%,在实际网络中,同步发生的概率更小,可以近似地认为真实网络中的同步现象不发生.Appenzeller 进一步指出,缓存中队列的分布也服从正态分布,其均值和标准偏差与 TCP 流的数量 N 以及缓存的大小 B 有关,标准偏差为

$$\delta_Q=\delta_w\leq\frac{1}{3\sqrt{3}}\frac{RTT\times C+B}{\sqrt{N}},$$

并得出链路利用率的下界表达式:

$$\text{Utilization}\ N\geq erf\left(\frac{3\sqrt{3}}{2\sqrt{2}}\frac{B}{\frac{RTT\times C+B}{\sqrt{N}}}\right).$$

在速率为 2.5Gb/s 或者 10Gb/s 的典型骨干链路上,小缓存法则可将经验法则的路由器缓存需求降低 2 个数量级,而且能够使路由器具有更好的时延特性.

- 小缓存法则的验证

文献[22]中已经对小缓存法则进行了仿真验证,而且文献[23]还给出了许多基于实际网络的近期实验结果.这些实际网络包括利用商用路由器搭建的实验网络和实际运行的主干网络.作者在不同的流量模式、网络拓扑、路由器结构和流量测量方法下进行了大量重复实验,在每一次独立实验中,缓存数量被降低不同的值,以便

于判定缓存数量在降低到多少时,链路利用率才开始降低.汇总的实验结果表明:当缓存数量大于 $RTT \times C/\sqrt{N}$ 时,链路利用率无损失发生,当缓存数量接近于 $RTT \times C/\sqrt{N}$ 时,链路利用率开始降低.

以其中的一个实验来说明,实验对象选择美国国内实际运行的骨干网络,其链路速率为 2.5Gb/s,属于拥塞链路,链路利用率每天至少有几个小时达到 90%以上(运营商正在考虑对链路进行扩容),网络中,Juniper 路由器的缺省缓存容量是 190ms,研究人员将缓存的值分别降为 10ms,5ms,2.5ms 和 1ms,发现当缓存值降低到 5ms 时,在实验周期(5~7 天)内未发现有任何数据包丢失,即使在链路利用率达到 95%时;而当缓存值降低到 2.5ms 和 1ms 时,开始有数据包被丢弃,但丢包率低于 0.2%.

- 关于小缓存法则的讨论

小缓存法则的成立基于两点假设:(1) 链路利用率是决定路由器缓存设置的指标;(2) 当链路上 TCP 流数目巨大时,TCP 是非同步传输的.

链路利用率是网络运营商敏感的指标,利用率达到 100%就意味着运营商能够最大化使用已投资的链路资源,从而能够取得最大的经济利益.而对终端的用户来说,链路利用率并不是其关心的指标,用户所关注的仅仅是更小的延迟和更低的丢包率.若小缓存法则成立,则小缓存在满足链路利用率的同时,显然还能有效地降低排队时延和延迟抖动,从而获得更好的时延特性;而对丢包率来说,情况就相对复杂一些:过高的丢包率将降低 TCP 流的吞吐率,而过低的丢包率将使得 TCP 反馈机制失效.

尽管仿真和实验都表明:当 TCP 流数目巨大时,TCP 是非同步传输的,但至目前为止,在这一结论上还未能形成一致的认识.为了理解 TCP 流数目和同步化之间的关系,Raina 和 Wischik 在文献[24]中对不同缓存值下的网络进行了模拟,认为:TCP 流的同步将导致聚合窗口的周期性变化,小缓存法则将导致网络不稳定,吞吐率降低.与此相对的是,Ganjali 等人的仿真却发现,小缓存法则虽然对网络的稳定性产生了一定的影响,但其影响仅在一定的限度内,并不像 Raina 和 Wischik 所认为的那样显著[23].导致这两种不同结论的原因可能在于,Raina 的数学模型中丢包率指标是公平性的,而实际中,当 TCP-Reno 和尾丢弃策略(drop-tail)共同实现时,丢包将是非公平性的*.

小缓存法则中,链路上 TCP 长流数目和短 TCP 流数目的比例如何?长 TCP 流数目 N 如何确定?这些问题到目前为止也无明确的结论.因此,关于小缓存法则比较慎重的结论应该是:骨干链路上 TCP 流数目巨大时,路由器缓存需求可至少降低至经验法则的 1/10,并且能够保证网络性能不因缓存数量的减少而产生降低,而且,随着缓存的减少,网络时延和时延抖动性能还能得到提高.

2.3 基于丢包率的缓存法则

- 仿真环境和结论

Dhamdhere 和 Dovrolis 通过对边缘网络中一条拥塞严重的低容量瓶颈链路的研究(如图 3 所示)发现,小缓存法则容易使丢包率明显上升,最高可达 17%,严重影响路由器的性能和网络的稳定.因此,他们提出了基于丢包率的缓存法则(drop-based rule),认为路由器中应该设置更大容量(在某些情况下甚至大于"经验法则"的缓存需求)的缓存[3].文献[25]通过仿真发现,丢包率的上升会引起吞吐量的下降和 TCP 流传输时延的增大,同时会引起带宽的分布不均匀.

Drop-Based rule 来源于仿真分析,图 3 是仿真的拓扑图,其中,目标链路的容量为 50Mb/s,共享的 TCP 流数目约为 200 条,其中大部分为长 TCP 流和极少的短 TCP 流.18 个源节点以平衡二叉树方式排列.由源节点生成的 TCP 流的 RTT 从 30ms~530ms 不等,有效 RTT[26](effective RTT)为 60ms.

- 基于丢包率法则的讨论

Drop-Based rule 表明,小缓存规则并不是一种普适的缓存设置规则,正如 Dhamdhere 和 Dovrolis 的测试结果所表现的那样,过小的缓存在边缘网络的适应性就不尽如人意.但单纯增大缓存容量也是不可行的:一方面,

* Wang,Ganjali. Unifying buffer sizing results through fairness. Manuscript submitted for publication. Also available as technical report, HR06-HPNG-060606, Stanford University, June 2006.

当链路发生拥塞时,我们希望“坏消息”能够尽快地传回发送端,从而在尽可能短的时间内减少发送端 TCP 的窗口大小,若缓存设置过大,则“坏消息”反馈至发送端所用的时间将大为增加;另一方面,若设置的缓存过小,那么其导致丢包率过高的后果会使得 TCP 的传输性能趋于崩溃.

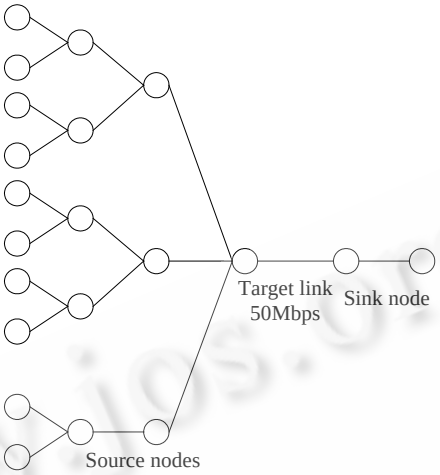

Fig.3 Experimental topology in drop-based buffer rule

图 3 基于丢包率的缓存分析拓扑

2.4 极小缓存法则

- 分析假设和结论

Enachescu 等人在文献[27]中提出了极小缓存法则(tiny buffer rule),认为若可以牺牲小量的吞吐率指标,核心路由器仅需几十个包的缓存即可满足要求,确切地说,$O(\log W)$缓存即可满足核心路由器需求,其中,W 代表拥塞窗口的值.Raina 和 Wischik 在文献[24]中也对极小缓存法则表示了一定的认同,认为极小缓存法则不仅分析了当前核心路由器缓存需求问题,也为未来全光路由器中的光缓存实现难题做了有益的探索.

极小缓存法则成立的关键在于要求接入链路的速率要远低于核心链路的速率,目的是能够将到达核心链路的聚合流量进行有效平滑,使其服从泊松分布,从而能够在牺牲一定链路利用率条件下将缓存需求降低至几十个包,同时保证丢包率满足需求.实际网络中,接入链路速率通常远低于核心链路速率(目前核心链路速率已达 10Gbp/s,而接入链路由于受到经济和其他原因的影响,速率远低于 10Gbp/s).此外,还有学者通过使用 Paced TCP[28]的方法获得同样的聚合流量平滑效果,从而满足极小缓存法则的要求.

- 极小缓存法则的验证

Ganjali 等人针对极小缓存法则在 Sprint ATL 展开了实验,Verizon 通信和 Lucent 也有相关研究小组进行了类似实验.在这些实验中,采用了商用的网络流量生成器和大量 Linux boxes 产生流量的方法来生成 1Gb/s 的长 TCP 流(主要为 FTP 和 HTTP),流量输入至极小缓存法则的路由器之后,利用路由器本身的统计功能或流量生成器提供的统计功能可对吞吐率、丢包率和时延等性能进行统计.统计结果显示:在满足上述极小缓存法则条件下,路由器的性能仅有微小的下降[23].

- 极小缓存法则的讨论

将光技术逐渐引入路由器设计,乃至最终实现全光路由器无疑是今后相当长时间内路由器发展的方向.在光技术引入路由器设计的道路上,有一个难题一直未能得到有效的解决,即光存储问题.虽然目前的研究已经可在一个集成光电芯片上实现容量为几十个分组的光 FCFS(first come first serve)缓存[29],但大规模光缓存的实现除了采用笨拙的光纤迂回策略外,目前还没有效方法.极小缓存法则能够将全光网络中充裕的带宽资源和复杂的缓存实现加以折衷,在牺牲小量带宽利用率(10%~15%)的条件下,使得几十个包的小缓存就能满足要求,这意

味着光缓存难题可以得到有效的缓解乃至消除.而且即使在当前的网络中,核心链路的链路利用率也仅为20%~30%[30,31],以牺牲 10%~15%的链路利用率来换取极小的路由器缓存也具有实际意义.

极小缓存法则是否能够真正指导核心路由器设计?其对网络性能的影响究竟如何?对链路利用率和丢包率的影响到底有多大?这些问题至今还未能形成定论,需要研究人员进一步地开展研究和实验.

2.5 小 结

基于 TCP 机制研究路由器的缓存需求已经成为当今网络领域的热点之一.一些著名的大学及科研机构如 Stanford 大学、Bell 实验室以及众多的路由器设备提供商等相继致力于这方面的研究,并取得了众多的研究成果,包括经验法则、小缓存法则、基于丢包率的缓存法则和极小缓存法则等,但是至今为止,不同法则的结论差别巨大,未能统一.综合各个结论可以发现,是不同的研究条件和不同的分析目标导致法则的结论产生矛盾,见表 1.

Table 1 Summary of buffer sizing rules

表 1 缓存法则小结

Rule name	Object	Assumptions	Buffer size	Performance
Rule-of-Thumb	100% link utilization	A single long-lived TCP	$RTT \times C$	Utilization: 100% Drop rate: Low Delay: Large
Small buffer rule	100% link utilization	N long-lived TCP	$RTT \times C / \sqrt{N}$	Utilization: 100% Drop rate: Up to 17% Delay: Small
Drop-Based rule	Drop rate	A heavily congested low capacity bottleneck link	Much larger buffers even than $RTT \times C$	Utilization: 100% Drop rate: Low Delay: Large
Tiny buffer rule	High link utilization	Access links are much slower than the core links, or Paced TCP	$O(\log W)$	Utilization: 80%~85% Drop rate: High Delay: Small

小规模的缓存的确能够维持 TCP 链路的高利用率,但往往无法很好地满足丢包率的性能要求;大规模的缓存能够满足丢包率和链路利用率要求,但又会使时延大为增加.因此,如何利用路由器缓存实现丢包率、链路利用率和时延等性能指标的最佳折衷,应该是后续研究的努力方向.

3 其他缓存研究结论

高速路由器是由链路输入/输出/复用、转发引擎、交换网络、系统监控等功能模块耦合后的复杂系统,很难对其建立准确的数学模型.以往的缓存分析基本上都采用简单的输出排队模型(output-queued),而在实际路由器中,输出排队由于需要很高的加速比,实现难度大,一般不被采用.当前,路由器大都采用联合输入/输出排队(combined input-output queued,简称 CIOQ)模型.Beheshti 等人基于 CIOQ 模型研究了路由器的缓存需求,认为:在加速比为 2 的条件下,几十个包的缓存可满足 CIOQ 路由器要求[32].

近年来,对 IP 网络的网络性能进行测量、分析、评价、调整受到越来越多的关注.基于测量技术分析路由器缓存需求就是在这种情况下出现的.测量显示,边缘网络的设备上容易发生网络的拥塞故障.据 MBone 的数据丢失模型研究显示,大多数的数据丢失事件都发生在网络的边缘,而并非是网络高速干线上的骨干节点.另外,对 Sprint 骨干节点路由器的测量研究发现[33],骨干路由器缓存队列的长度将很少能够超过 10 个包,产生这种现象的主要原因可能是 Sprint 网络中链路的利用率通常低于 20%的缘故.文献[34]对实际网络中运行的路由器的测试也同样表明,骨干网络中路由器的缓存需求远远小于其目前的设置值.然而到目前为之,测量技术仅能够确定路由器的缓存需求小于路由器现今所设置的缓存值,还未能给出更加具体的结论.

目前,在路由器缓存需求分析还无定论的情况下,各路由器厂商仍然普遍采用传统的经验法则来指导商用路由器设计,如何实现高速、大容量数据包缓存就成为高性能网络设备设计的瓶颈[35,36].例如,当链路速率为 40Gb/s 时,对 40 字节的数据包来说,要求存储器的随机存取时间(random access time)不大于 4ns;若 TCP 连接的平均往返时间选为 250ms,则依据经验法则要求存储器的容量满足 10Gbits(1.25Gbytes),在目前的存储器工艺水

平下,如此高速、大容量的数据包存储器很难实现.另外,从图 1 可以发现,目前动态随机存储器(dynamic random access memory,简称 DRAM)的带宽每 18 个月增长 1.1 倍,而骨干网带宽每 7 个月翻一番,因此,存储带宽的瓶颈效应将会越来越严重[37,38].Stanford 大学的 Iyer 等人利用 DRAM 的大容量和随机静态存储器(SRAM)高速的特点,提出混合结构(SRAM 与 DRAM)联合工作模式[39],在一定程度上缓解了数据包存储器的容量和访问速度问题,但是其致命缺陷在于该结构受限于 SRAM 的访问速度.即如果线路速率超过 SRAM 所能提供的带宽,这种 SRAM 与 DRAM 混合结构就会因 SRAM 的速度不够而彻底失去作用.此外,文献[40]提出了一种新型的三级存储阵列结构,可以成功地解决数据包存储器的容量和带宽问题,理论上可以实现任意高速数据包的缓存.

4 结 论

路由器是一种存储转发设备,其内部的缓存一方面能够提高链路使用率,减少路由器丢包数,另一方面,路由器缓存也能够增加网络的时延和时延抖动,进而降低整个网络性能.因此,路由器的缓存分析已经成为当今网络领域的热点问题.

路由器缓存需求的分析方法可以概括为两类:一类是利用经典的排队理论对路由器的缓存需求进行分析,在给定的输入流量模型的条件下,推导出缓存和丢包率、时延等性能指标间的对应关系.该类方法的分析结论依赖于流量的输入模型,在到达报文符合短相关输入模型时,小缓存便可满足路由器的性能指标要求,而当到达报文符合长相关输入模型时,路由器需要大容量的缓存才能满足性能指标要求;另一类方法以 TCP 协议模型为基础进行分析,并取得了众多的研究成果,包括经验法则、小缓存法则、基于丢包率的缓存法则和极小缓存法则等.由于各种法则在分析过程中的分析条件和分析目标各不相同,导致了不同法则的分析结论差别巨大,未能统一.总的来说,小规模的缓存的确能够维持 TCP 链路的高利用率,但往往无法很好的满足丢包率的性能要求;大规模的缓存能够满足丢包率和链路利用率要求,但又会使时延大为增加.因此,如何利用路由器缓存实现丢包率、链路利用率和时延等性能指标的最佳折衷,应该是后续研究的方向.

References:

[1] Bradner S. Benchmarking terminology for network interconnection devices. RFC 1242, 1991. http://www.ietf.org/rfc/rfc1242.txt

[2] Bradner S, McQuaid J. Benchmarking methodology for network interconnect devices. RFC 2544, 1999. http://www.ietf.org/rfc/rfc2544.txt

[3] Dhamdhere A, Dovrolis C. Open issues in router buffer sizing. ACM SIGCOMM Computer Communications Review, 2006,36(1): 87−92.

[4] Wischik D. Buffer requirements for high-speed routers. In: Proc. of the ECOC 2005. 2005. 23−26. http://www.cs.ucl.ac.uk/staff/D.Wischik/Research/

[5] Zhao Z, Darbha S, Reddy ALN. A method for estimating the proportion of nonresponsive traffic at a route. IEEE/ACM Trans. on Networking, 2004,12(4):708−718.

[6] Erramilli A, Narayan O, Willinger W. Experimental queueing analysis with long-range dependent packet traffic. IEEE/ACM Trans. on Networking, 1996,4(2):209−223.

[7] Leland WE, Taqqu MS, Willinger W, Wilson DV. On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Trans. on Networking, 1994,2(1):1−15.

[8] Jin ZG, Y JS, Wang JD, Shu YT. Sampling of important events in performance evaluation of networks. Journal of Tianjin University, 2001,34(6): 738−740 (in Chinese with English abstract).

[9] Huang C, *et al*. Fast simulation of self-similar traffic in ATM networks. In: Proc. of the ICC'95. 1995. 438−444.

[10] Neidhardt AL, Wang JL. The concept of relevant time scales and its application to queuing analysis of self-similar traffic (or is hurst naughty or nice?). In: Proc. of the ACM SIGMETRICS 1998. 1998. 222−232.

[11] Likhanov N, Tsybakov B, Georganas ND. Analysis of an ATM buffer with self-similar ("Fractal") input traffic. In: Proc. of the IEEE INFOCOM'95. Washington: IEEE Computer Society,1995. 985−992.

[12] Norros I. On the use of fractional Brownian motion in the theory of connectionless networks. IEEE Journal on Selected Areas in Communication, 1995,13(6):953−962.

[13] Norros I. A storage model with self-similar input. Queueing Systems, 1994,16:387−396.

[14] Duffield ND, O'Connell N. Large deviations and overflow probabilities for the general single-server queue, with application. In: Proc. of the Cambridge Philosophy Society. 1995. 363−375. http://www.research.att.com/~duffield/papers/

[15] Rananand N. Upper-Bounds for tail probability of a queue with Long-range dependent input. In: Proc. of the ICC'98. IEEE Press, 1998. 1466−1472.

[16] Erramilli A, Singh PR, Pruthi P. An approach of deterministic chaotic maps to model packet traffic. Queuing System, 1995,17: 171−206.

[17] Pruthi P, Erramilli A. Heavy-Tailed ON/OFF source behavior and self-similar traffic. In: Proc. of the ICC'95. 1995. 445−450.

[18] Rao YH. Self-Similar network traffic and performance of high-speed routing structure [Ph.D. Thesis]. Wuhan: Huazhong University of Science and Technology, 2004 (in Chinese with English abstract).

[19] Villamizar C, Song C. High performance TCP in ANSNET. ACM Computer Communication Review, 1994,24(5):45−60.

[20] Bush R, Meyer D. Some Internet architectural guidelines and philosophy. RFC 3439, 2002. http://www.rfc-archive.org/getrfc.php?rfc=3439

[21] Fraleigh CJ. Provisioning Internet backbone networks to support latency sensitive applications [Ph.D. Thesis]. Palo Alto: Stanford University, 2002.

[22] Appenzeller G, Keslassy I, McKeown N. Sizing router buffers. In: Proc. of the SIGCOMM 2004. New York: ACM Press, 2004. 281−292.

[23] Ganjali Y, McKeown N. Update on buffer sizing in Internet routers. Computer Communications Review, 2006,36(5):67−70.

[24] Raina G, Wischik D. Buffer sizes for large multiplexers: TCP queueing theory and instability analysis. In: Proc. of the EuroNGI. Rome, 2005. 173−180. http://www.cs.ucl.ac.uk/staff/d.wischik/Research/tcptheory.html.

[25] Morris R. Scalable TCP congestion control. In: Proc. of the IEEE INFOCOM 2000. 2000. 1176−1183.

[26] Dhamdhere, Jiang H, Dovrolis C. Buffering sizing for congested Internet links. In: Proc. of the IEEE INFOCOM 2005. 2005. 1072−1083. http://www.cc.gatech.edu/~amogh/buffers.pdf

[27] Enachescu M, Ganjali Y, Goel A, McKeown N, Roughgarden T. Routers with very small buffers. In: Proc. of the IEEE INFOCOM 2006. 2006.

[28] Aggarwal A, Savage S, Anderson T. Understanding the performance of TCP pacing. In: Proc. of the IEEE INFOCOM. 2000. 1157−1165. http://www.cs.ucsd.edu/~savage/papers/Infocom2000pacing.pdf

[29] Park H, Burmeister EF, Bjorlin S, *et al*. 40-Gb/s optical buffer design and simulations. In: Proc. of the Numerical Simulation of Optoelectonic Devices (NUSOD). 2004. 9−20.

[30] Odlyzko AM. The current state and likely evolution of the Internet. In: Proc. of the IEEE GLOBECOM'99. IEEE, 1999. 1869−1875.

[31] Odlyzko AM. Data networks are lightly utilized, and will stay that way. Review of Network Economics, 2003,2(3):210−237.

[32] Beheshti N, Ganjali Y, Rajaduray R, Blumenthal D, McKeown N. Buffer sizing in all-optical packet switches. In: Proc. of the OFC/NFOEC. 2006. http://yuba.stanford.edu/buffersizing/

[33] Hohn N, Veitch D, Papagiannaki K, Diot C. Bridging router performance and queuing theory. In: Proc. of the SIGMETRICS. New York: ACM Press, 2004. 355−366.

[34] Appenzeller G. Sizing router buffers [Ph.D. Thesis]. Stanford University, 2004.

[35] Shah D, Iyer S, Prabhakar B, McKeown N. Maintaining statistics counters in router line cards. IEEE Micro, 2002,22(1):76−81.

[36] Iyer S, Zhang R, McKeown N. Routers with a single stage of buffering. In: Proc. of the ACM SIGCOMM. Pittsburgh, 2002. Computer Communication Review, 2002. 431−439.

[37] Iyer S, Kompella RR, McKeown N. Techniques for fast packet buffers. Technical Report, TR01-HPNG-081501, Stanford University, 2001.

[38] Sailesh K, Venkatesh R, Philip J, Shukla S. Implementing parallel packet buffering: Part 1. In: Proc. of the Commsdesign. 2002. http://www.commsdesign.com/story/OEG20020422S0006

[39] Iyer S, Kompella RR, McKeown N. Analysis of a memory architecture for fast packet buffers. In: Proc. of the IEEE High Performance Switching and Routing. 2001. 368−373.

[40] Wang P, Yi P, Jin DP, Zeng LG. Buffering high-speed packets with tri-stage memory array and its performance analysis. Journal of Software, 2005,16(12):2181−2189 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/2181.htm

附中文参考文献:

[8] 金志刚,杨晋生,王继东,舒炎泰.网络性能评价中重要事件采样仿真.天津大学学报,2001,34(6):738−740.

[18] 饶云华.自相似网络通信量及高速路由结构性能研究[博士学位论文].武汉:华中科技大学,2004.

[40] 王鹏,伊鹏,金德鹏,曾烈光.基于三级存储阵列缓存高速数据包及性能分析.软件学报,2005,16(12):2181−2189. http://www.jos.org.cn/1000-9825/16/2181.htm

李玉峰(1976－),男,山东烟台人,博士生,主要研究领域为宽带信息网络,高速路由器核心技术.

邱菡(1981－),女,博士生,主要研究领域为宽带信息网络,流媒体技术.

兰巨龙(1962－),男,教授,博士生导师,主要研究领域为宽带信息网络.

汪斌强(1963－),男,教授,博士生导师,主要研究领域为宽带信息网络.