ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.6, June 2008, pp.1499−1507
DOI: 10.3724/SP.J.1001.2008.01499

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

TCP 流竞争拥塞及拥塞链路的缓存需求研究*

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)

Study on TCP Flow-Competing Congestion and Buffer Requirement of the Congested Links

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

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

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

+ Corresponding author: E-mail: lyf@mail.ndsc.com.cn

Li YF, Qiu H, Lan JL, Wang BQ. Study on TCP flow-competing congestion and buffer requirement of the congested links. *Journal of Software*, 2008,19(6):1499−1507. http://www.jos.org.cn/1000-9825/19/1499.htm

Abstract: This paper first presents an analysis model named flow-competing congestion model (FCCM) for this type of congestion. Based on FCCM, the paper derives the distribution of competing flows at the congested link, and analyzes the conditions under which the flow-competing congestion would happen. This paper also explores how much buffers a congested link requires to keep full link utilization when the flow-competing congestion occurs. This paper proves that when sizing buffers for a congested internet link with the aim of keeping full link utilization, the buffer requirement of the flow-competing congestion is not bigger than the minimum buffer requirement of the famous BSCL (buffer sizing for congested internet links) scheme.

Key words: TCP flow; flow-competing congestion; buffer requirement; competing flow

摘 要: 针对流竞争拥塞,提出了一种拥塞分析模型 FCCM(flow-competing congestion model),给出了 TCP 竞争流在拥塞链路上的分布特性,推导了流竞争拥塞发生的条件,进而分析了在流竞争拥塞发生时,路由器为维持拥塞链路100%利用率所需的最小缓存.分析结果表明,当流数目不确定时,应对流竞争拥塞所需的缓存将不大于流数目确定时经典 BSCL(buffer sizing for congested internet links)方案中的最小缓存需求.

关键词: TCP 流;流竞争拥塞;缓存需求;竞争流

中图法分类号: TP393 **文献标识码**: A

路由器是一种存储转发设备,其内部的缓存是分组网络的重要组成部分.一方面,设置大容量的路由器缓存能够更好地吸收链路上突发性变速率到达的业务,当链路上发生拥塞时,能够对新进入的数据包进行暂存,降低丢包率,维持高链路利用率;另一方面,过大的路由器缓存将导致数据包排队时延的增大,引起时延抖动,降低 TCP(transmission control protocol)流的吞吐率;此外,小缓存的路由器还能推动全光路由器的建设,降低了路由

* 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 2006-11-20; Accepted 2007-02-08

器的设计复杂度,增加了路由器的可扩展性[1−3].因此,研究路由器所需的缓存容量不仅能够优化路由器的设计,提高路由器的性能,提高整个网络的性能,还将对于全光路由器设计中光缓存难题的解决具有重要意义.然而到目前为止,关于路由器的缓存容量需求问题仍未有科学的结论[4,5].

1994 年,Villamizar 和 Song 在文献[6]中给出了路由器缓存需求的经典结论,即著名的带宽时延乘积规则(bandwidth-delay product,简称 BDP).多年以来,这个规则一直指导着路由器的设计.BDP 规则就是:为了提高拥塞链路的利用率,路由器所需的缓存容量应当满足 $B=\overline{RTT}\times C$.其中,B 为路由器所需的缓存,$\overline{RTT}$ 是一个 TCP 连接的平均往返时间(round trip time),C 为拥塞链路的带宽.

Guido Appenzeller 等人在 2004 年提出了一种缓存分析模型——斯坦佛模型.斯坦佛模型的关键思想在于:当拥塞链路上 TCP 流的数目足够大且非同步时,路由器仅需少量的缓存便可追求链路的高利用率.其中,文献[7,8,10]认为:骨干网上核心路由器的缓存仅需满足 $B=\overline{RTT}\times C/\sqrt{N}$ 即可维持 100%的链路利用率,从而将 BDP 规则的缓存需求大为降低,其中,N 代表拥塞链路中长 TCP 流的数目.文献[7]中定义的那些从未离开过慢启动阶段的 TCP 流为短 TCP 流,反之则为长 TCP 流.更进一步地,Enachescu 等人在文献[9]中提出:在牺牲少量链路利用率的条件下,少于 20 个包的缓存容量就能够保证链路实现高吞吐率.

斯坦佛模型对路由器缓存需求的分析是基于链路利用率单个分析目标进行的.与斯坦佛模型不同,文献[11]探讨了在维持 100%链路利用率并且满足确定的丢包率上限值的双重目标上,路由器为应对链路拥塞所需的最小缓存值,也就是路由器缓存分析中著名的 BSCL(buffer sizing for congested internet links)方案.

假设 W 是拥塞发生前一刻 TCP 流的平均窗口值,明显地,当 $NW/\overline{RTT}\geq C$ 满足时,链路将发生拥塞.设 B 为拥塞链路的缓存需求,则为了应对拥塞应满足 $B=NW/\overline{RTT}-C$.不失一般性,假设拥塞链路的带宽 C 和平均往返时间 $\overline{RTT}$ 为定值,则流数目 N 和窗口值 W 就成为影响 B 值的关键因素.概括而言,斯坦佛模型和 BSCL 方案在分析路由器缓存需求的过程中,都是以长的 TCP 流为研究对象,假定拥塞链路上存在 N 个长 TCP 流并且这些 TCP 流永久存在,在此条件下,链路上拥塞的发生将仅由窗口值 W 的增长而导致.实际上,每一个 TCP 流都有一个有限的传输持续时间和长度,拥塞链路上 TCP 流的数目 N 也并不固定,而是一个变量.尤其需要指出的是,N 的突发性增大本身就能导致拥塞的发生.本文中,我们称这种由于目的输出链路相同的 TCP 流的数目的突发性增大超过了目的输出链路的处理能力而引起拥塞为“流竞争拥塞”.那么,流竞争拥塞发生的条件是什么?路由器为应对流竞争拥塞应设置多大的缓存?斯坦佛模型和 BSCL 方案由于所假定的研究条件的局限性,不可能对这些问题给出答案.本文中,我们给出了一种新的分析模型——FCCM(flow-competing congestion model),借助该模型,我们推导得出如下 3 点结论:

1) 当输入链路上的 TCP 流数目 N 和路由器的输入端口数目 M 都足够大时,若 $N/M\leq 5$,则目标链路上聚合后的 TCP 竞争流的流数目服从泊松分布;若 $N/M>5$,则服从正态分布.
2) 当路由器的输入端口数目 M 足够大时,目标链路上由 M 个输入链路聚合后的 TCP 竞争流的流数目分布与 M 无关.
3) 在维持目标链路 100%链路利用率条件下,应对目标链路上流竞争拥塞所需的最小缓存 B_{FCCM} 满足 $B_{FCCM}<B_{BSCL}$,其中,B_{BSCL} 为 BSCL 方案中的最小缓存需求.

从而为上述问题给出了完整的答案.

1 定义和模型

概括来说,以往对路由器缓存需求的分析都是基于如图 1 所示的模型进行的,斯坦佛模型和 BSCL 方案也是如此.在该模型中,$s_1,s_2,\ldots,s_N$ 代表 N 个 TCP 流的源端,$d_1,d_2,\ldots,d_N$ 代表对应 TCP 流的目的端,拥塞链路(v,w)的带宽为 C,所承载的长 TCP 流数目为 N,节点 v 通常被假设为一个基于输出交换(output queued,简称 OQ)[12]的路由器,其缓存大小为 B.在该模型中,由于 TCP 流的数目 N 为定值,链路(v,w)拥塞就由 TCP 流的平均窗口值 W 的增长导致流量超过 C 而引起.

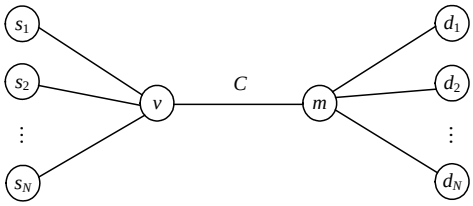

Fig.1 The model in most previous buffer sizing works

图 1 以往的路由器缓存分析模型

实际上,在具体的路由器输出链路上,除了上述的基于窗口 W 增长所导致的拥塞以外,常见的另一种拥塞是由于相同目的输出链路的 TCP 流争用而引发的,而且这种拥塞无法通过路由器交换结构和调度算法的优化设计来避免,即便是在提供服务质量保障方面最具性能优势的 OQ 交换结构也无能为力.以一个 4×4 交换结构为例,如图 2 所示,一旦 4 个输入链路上的 TCP 流同时流向相同的输出链路时,仅有其中 1 个输入链路上的 TCP 流的报文可以获得调度输出,其余输入链路上的 TCP 流的报文被阻塞.本文中,我们称这种由于目的输出链路相同的 TCP 流数目的突发性增大超过了目的输出链路的处理能力而引起的拥塞为“流竞争拥塞”.

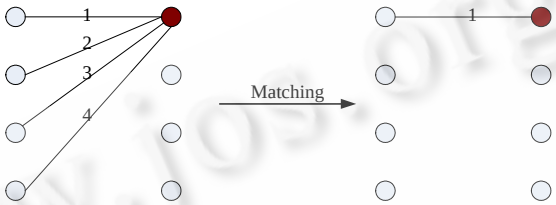

Fig.2 An example of congestion resulted from flows competing for a same link

图 2 TCP 流竞争同一链路导致拥塞的举例图

针对流竞争拥塞,我们做如下的定义:

定义 1. 任选路由器的一条输出链路为研究对象,则该输出链路即称为目标链路.

定义 2. 路由器所有输入链路上,在同一时刻目的输出链路都为目标链路的 TCP 流,称为竞争流.

为进一步分析流竞争拥塞,我们结合现代路由器实现结构[15]给出了如图 3 所示的流竞争拥塞模型 FCCM.在 FCCM 中,假设交换结构为 OQ,并且能够达到 100%吞吐率.以 $l_i(1\le i\le M)$代表路由器的输入链路,$L_i(1\le i\le M)$代表路由器的输出链路,并且选择 L_i 作为目标链路.

为分析方便,我们假设每一条输入链路上所承载的长 TCP 流数目都为 N,假设 $l_i(1\le i\le M)$的链路利用率统一为ρ,由后面的分析可知,若假设每条输入链路的利用率都不相同,则本文所得出的结论仍将成立.此外,用 $S_{i,j}(1\le i\le M,1\le j\le N)$表示 l_i 上的第 j 个 TCP 流,以 $T_j(1\le j\le N)$表示 $S_{i,j}$ 的平均往返时间.同样地,用 $D_{i,k}(1\le i\le M,1\le k\le K_i)$表示 L_i 上的第 k 个 TCP 流,其中,K_i 表示链路 L_i 上总的 TCP 流数目.由此,我们得到 $MN=\sum_{i=1}^{M}K_i$.

需要说明的是,FCCM 与以往的缓存分析模型在本质上是不相同的:一方面,在以往的缓存分析模型中,拥塞链路上假定存在 N 个长 TCP 流,并且这些 TCP 流永久存在.而在 FCCM 中,目标链路上 TCP 流的数目是一个由 M 个输入链路聚合后的结果,其值是一个变量,变化范围为$[0,MN]$;另一方面,如前所述,以往的缓存分析模型并不能反映出流竞争拥塞这种拥塞类型.而 FCCM 模型由于包含有交换结构,因此能够反映流竞争拥塞.

路由和交换设备需要设置恰当的缓存来应对流竞争拥塞,否则,当流竞争拥塞发生时,将直接导致路由器的时延、时延抖动以及丢包率等方面的性能恶化,而且一旦发生大量丢包必将严重影响到网络的性能.如本文第 1 节所述,不同的缓存设置目的将导致不同的缓存需求分析结论.下面我们将以维持目标链路 100%利用率为目的,研究应对流竞争拥塞所需的最佳缓存值.首先分析目标链路上 TCP 竞争流的分布.

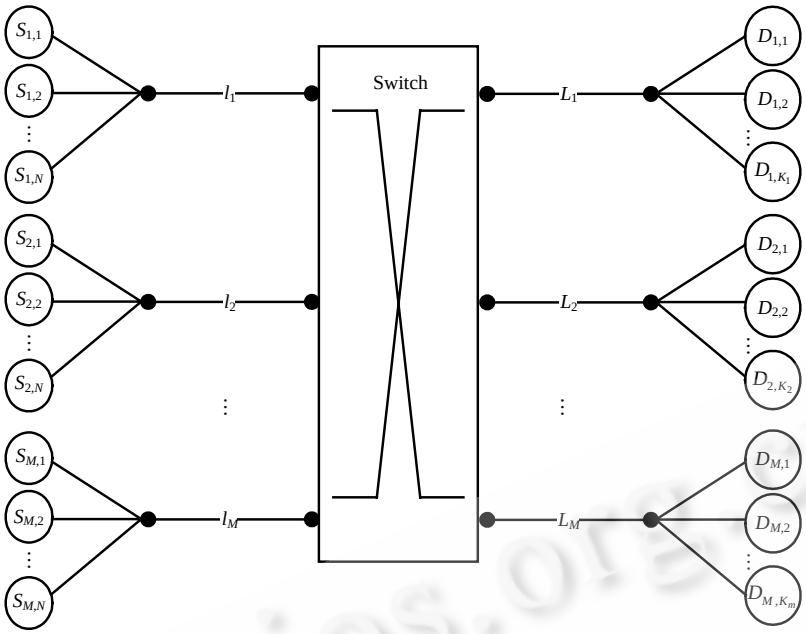

Fig.3 Flow-Competing congestion model

图 3 流竞争拥塞分析模型

2 目标链路上 TCP 竞争流的分布

从长期统计平均的角度来看,输入链路 l_i 的 TCP 流在 M 个目的输出链路上将服从均匀分布.因此,对 l_i 上的一个 TCP 流而言,其去往目标链路 L_i 的概率 P 为 $1/M$.在分析中,考虑到目的地址为本路由器的 TCP 流在 l_i 上所占的比重极小,我们对此做了忽略.此外,我们还假设 l_i 上的各个 TCP 流都是独立的,这一点显然也能够成立.由此,下述定理 1 将成立.

定理 1. 当输入链路上的 TCP 流数目 N 和路由器的输入端口数目 M 都足够大时,若 $N/M\leq 5$,则目标链路上聚合后的 TCP 竞争流的流数目服从泊松分布;若 $N/M>5$,则服从正态分布.

证明:以 X_i 表示输入链路为 l_i 并且目的输出链路为 L_i 的 TCP 流的数目,其取值范围为{0,1,…,N}.由于 l_i 上的各个 TCP 流都是独立的,并且在 M 个输出链路上服从均匀分布,因此,$X_i\sim Bin(N,P)$,具体为

$$P\{X_i=k\}=\binom{N}{k}P^k q^{n-k},$$

其中,$P=1/M$,$q=1-P$,X_i 的均值和方差为

$$E(X_i)=NP, Var(X_i)=NP(1-P).$$

若我们对上式取极限,$N\to\infty$,$P\to 0$,并且取 $NP=\lambda$,则将得到

$$\lim_{\left\{\begin{matrix}P\to 0\\ N\to\infty\\ NP\to\lambda\end{matrix}\right\}}\binom{N}{k}P^k q^{N-k}=\mathrm{e}^{-\lambda}\frac{\lambda^k}{k!} \tag{1}$$

式(1)表明,若 N 和 M 足够大,以至于满足条件 $N\to\infty$,$P=1/M\to 0$ 和 $NP\to\lambda$ 时,X_i 可认为服从泊松分布.经验上,当 $NP\leq 5$ 时,用式(1)取得的计算效果比较理想,而当 $NP>5$ 时,二项分布取正态分布为极限分布将取得更好的计算效果[13].总之,

$$X_i \sim \begin{cases} \pi(NP), & N/M \le 5 \\ Normal(NP, NP(1-P)), & N/M > 5 \end{cases}.$$

若以 X 表示目标链路 L_i 上的 TCP 竞争流数目,则有 $X = \sum_{i=1}^{M} X_i$.

显然,路由器各个输入链路间相互独立,即 $X_i(i=1,2,\ldots,M)$相互独立.从而有式(2)成立.

$$X \sim \begin{cases} \pi(MNP), & N/M \le 5 \\ Normal(MNP, MNP(1-P)), & N/M > 5 \end{cases} \tag{2}$$

定理 1 得证. □

网络传输速率和业务的快速增长促进了路由器端口速率的快速提高,目前,高端路由器的端口速率通常都达到了 10Gb/s.如此高的端口速率决定了输入链路上 TCP 流的数目 N 必然是一个较大的值.在此基础上,定理 1 表明,若路由器端口数目 M 巨大,则路由器的大容量端口交换结构本身就使得目标链路上的 TCP 流服从泊松分布或者正态分布.

定理 2. 当路由器的输入端口数目 M 足够大时,目标链路上由 M 个输入链路聚合后的 TCP 竞争流的流数目分布与 M 无关.

证明:当 P 满足 $P=1/M\to 0$ 时,式(2)可以变换成

$$X \sim \begin{cases} \pi(N), & N/M \le 5 \\ Normal(N, N), & N/M > 5 \end{cases} \tag{3}$$

式(3)表明,当 M 足够大时,目标链路上聚合后的 TCP 竞争流的数目将与 M 无关,定理 2 成立.同时,式(3)还表明,流竞争拥塞将仅与输入链路上承载的 TCP 流数目 N 有关,直观上理解,路由器的输入链路数目 M 越大,流竞争拥塞将越严重的观点并不成立.实际上,从长期统计平均上来看,只有每一条输入链路上的 TCP 流数目 N 越大,目标链路上聚合后的 TCP 竞争流数目的均值和方差才越大,流竞争拥塞的可能性也才越大. □

3 流竞争拥塞发生时目标链路的缓存需求分析

本节将以维持目标链路的 100%利用率为目的,研究应对流竞争拥塞所需的最佳缓存值.首先分析流竞争拥塞发生的条件.

3.1 流竞争拥塞发生的条件

引理 1. 当 $N>9, M>1$ 且输入链路的链路利用率满足 $1 \ge \rho \ge \frac{N}{N+3\sqrt{N}}$ 时,目标链路上可能发生流竞争拥塞.

证明:在实际的高端路由器中,输入链路上的 TCP 流数目 N 和路由器的输入端口数目 M 基本上都能自然满足 $N/M>5$ 和 $P=1/M\to 0$ 这两个条件,因此在下面的分析中,我们将选择 $X\sim Normal(N,N), 0\le X\le MN$,并以 $f_X(x)$作为 X 的概率密度函数,从而可以得到:

$$\int_0^{MN} f_X(x)\mathrm{d}x = 1.$$

若以 w'代表能够恰好充满目标链路所需的 TCP 流数目,近似地,我们可以得到:

$$w' = \frac{N}{\rho} \tag{4}$$

设$\beta=Pro${目标链路上发生流竞争拥塞},则有

$$\beta = 1 - \int_0^{w'} f(x)\mathrm{d}x,\ 0 \le x \le w'.$$

若引入新的虚变量 Y 满足 $Y\sim Normal(N,N), -\infty\le Y\le+\infty$,则其概率密度函数为

$$f(y) = \frac{1}{\sqrt{2\pi N}} \mathrm{e}^{-\frac{(y-N)^2}{2N}},\ -\infty < y < +\infty.$$

从而有

$$Z=\frac{Y-N}{\sqrt{N}}\sim Normal(0,1).$$

由标准正态分布的特性可知,当–3<*Z*<3 时,将有 $\int_{-3}^{3} f(z)\mathrm{d}z\cong 1$ 成立,从而可得:

$$\int_{N-3\sqrt{N}}^{N+3\sqrt{N}} f(y)\mathrm{d}y\cong 1,\ N-3\sqrt{N}<Y<N+3\sqrt{N}.$$

易推导得出,当 *N*>9 且 *M*>1 成立时,$0<N-3\sqrt{N}<Y<N+3\sqrt{N}<MN$ 成立.

比较变量 *X* 和虚变量 *Y* 可知,当 *N*>9 且 *M*>1 成立时,*f*(*x*)可相应地定义为

$$f(x)\cong\begin{cases}\dfrac{1}{\sqrt{2\pi N}}\mathrm{e}^{-\frac{(x-N)^2}{2N}}, & N-3\sqrt{N}<X<N+3\sqrt{N}\\ 0, & \text{else}\end{cases}\tag{5}$$

此时,β可相应地定义为

$$\beta=1-\int_0^{w'}\frac{1}{\sqrt{2\pi N}}\mathrm{e}^{-\frac{(x-N)^2}{2N}}\mathrm{d}x,\ N-3\sqrt{N}<x<N+3\sqrt{N}\tag{6}$$

w'代表的是恰好能够充满目标链路的 TCP 流数目,因此有

$$N-3\sqrt{N}<w'\tag{7}$$

式(5)和式(6)表明,目标链路若要能够发生流竞争拥塞,即要满足β>0,w'应满足:

$$w'<N+3\sqrt{N}\tag{8}$$

结合式(4)、式(7)和式(8),可得:

$$\frac{N}{N+3\sqrt{N}}\le\rho\le 1\tag{9}$$

总之,当 *N*>9 且 *M*>1 成立,且式(9)满足时β>0,引理 1 得证. □

此外,还可推导:

$$\beta=1-\frac{1}{2}\mathrm{erf}\left(\frac{N/\rho-N}{\sqrt{2N}}\right)-\frac{1}{2}\mathrm{erf}\left(\sqrt{\frac{N}{2}}\right).$$

上述 3 个条件中,*N*>9 且 *M*>1 在路由器中可自然成立,我们着重对式(9)的条件进行分析.首先,该条件表明,只有当输入链路的链路利用率足够大以至于满足式(9)时,目标链路上才有可能发生流竞争拥塞.其次,该条件还表明,当路由器输入链路的速率降低时,即输入链路上所能承载的 TCP 流数目 *N* 减少时,目标链路上将更容易发生流竞争拥塞;反之,当路由器输入链路的速率升高时,即目标链路上所能承载的 TCP 流数目 *N* 增大时,目标链路上将愈难发生流竞争拥塞.图 4 给出了目标链路上发生流竞争拥塞的条件曲线.明显地,当 *N*<100 时,输入链路的利用率满足ρ>0.77 就可能引起目标链路上流竞争拥塞的发生;而当 *N*>1000 时,输入链路的链路利用率只有满足ρ>0.92 才有可能导致流竞争拥塞的发生.对现有网络大量的测量显示,大多数的数据丢失事件都发生在网络的边缘,并非是网络的高速干线上的骨干节点,而现有网络发生数据包丢失事件的主要原因是网络拥塞事件.因此可以断定,网络的拥塞事件主要发生在接入链路而非核心链路.式(9)所示的条件为这一拥塞现象提供了有力的证据.

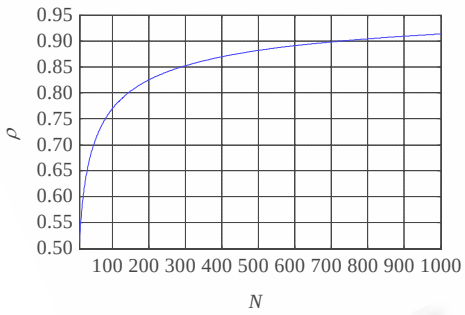

Fig.4 The situation line of flow-competing congestion happening

图 4 流竞争拥塞的条件曲线

3.2 流竞争拥塞发生时的缓存需求分析

以维持 100%链路利用率为目的,当目标链路上发生流竞争拥塞时,需要多大的缓存才能应对拥塞?本节给出定理 3 可作为答案.

定理 3. 在维持目标链路 100%链路利用率的条件下,应对目标链路上流竞争拥塞所需的最小缓存 B_{FCCM} 满足 $B_{FCCM}<B_{BSCL}$,其中,B_{BSCL} 为 BSCL 方案中的最小缓存需求.

证明:目标链路上发生流竞争拥塞时,拥塞将可能导致丢包,使得排队时延增大.目标链路上的 TCP 流感知到拥塞时,根据 TCP 拥塞窗口的 AIMD(additive increase multiplicative decrease)[14]算法,将加倍递减(multiplicative decrease)发送速率,将拥塞窗口大小减半,以适应目标链路的容量;而当拥塞消除后,根据 AIMD 算法,将逐步递增(additive increase)发送速率,每当收到一个确认 ACK(ACKnowledgement),TCP 就将拥塞窗口增大一个 S (segment size).在此情形下,为维持目标链路 100%链路利用率,必然要求缓存在目标链路因拥塞导致丢包前能够暂存足够的流量,以便当 TCP 流加倍递减发送速率时流量仍然能够充满目标链路.

假设流竞争拥塞发生前一刻,目标链路上竞争流的数目为 $x(w'\le x\le MN)$.此外,假设 $\overline{L}_x$ $(w'\le x\le MN)$为流竞争拥塞所导致的突发性丢包的平均长度.任选一个 TCP 流,则 $\overline{L}_x$ 个丢包中无一属于该 TCP 流的概率为 $(1-1/x)^{\overline{L}_x}$,因此,所有 x 个 TCP 流中至少丢失过一个数据包的 TCP 流所占的比例为

$$\theta=1-\left(1-\frac{1}{x}\right)^{\overline{L}_x}.$$

假设引理 1 的流竞争拥塞条件都能满足,拥塞发生前 TCP 流的平均窗口值为 W,则有下式成立:

$$xW\ge C\times RTT+B, w'<x\le MN.$$

由于链路上的 TCP 流是非全局同步[8]的,因此拥塞产生后,将仅有部分的 TCP 流产生丢包,这些 TCP 流的窗口将减少一半;而未产生丢包的 TCP 流将继续增大窗口值,增大的幅度为一个 S.设拥塞结束后的平均窗口值为 W_c,则拥塞结束后有

$$W_c=\theta\frac{W}{2}+(1-\theta)(W+S).$$

拥塞发生前一刻,目标链路和缓存都处于充满状态,$xW\ge C\times RTT+B$.拥塞发生后,要求缓存仍然能够维持链路的充满状态.由于本文考虑的是最小缓存需求,因此,我们可以考虑拥塞结束后的缓存为空的状态,此时有 $xW_c=C\times RTT$,结合上式可得:

$$\theta\frac{C\times RTT+B}{2x}+(1-\theta)\left(\frac{C\times RTT+B}{x}+S\right)=\frac{C\times RTT}{x}.$$

变换后求 B 可得:

$$B=\frac{\theta\times C\times RTT-2Sx(1-\theta)}{2-\theta},\ w'<x\le MN \tag{10}$$

显然,不同的 θ,C,RTT,S 和 x 将对应不同的 B.需要指出的是,x 是一个变量,代表的是流竞争拥塞发生时的竞争流数目,其取值范围为 $[w',MN]$.因此,式(10)中的缓存需求 B 也是一个变量,需要进一步讨论其最小值.

首先分析 B 的概率密度函数 $f_B(b)$,为简化分析,我们假设 $\rho=1$,不难看出,若选择 $\rho<1$,我们将得到一个更小的缓存需求值.

以 $F_B(b)$ 表示 B 的概率分布函数,可得:

$$\begin{aligned}F_B(b)=P\{B\le b\}&=P\left\{\frac{\theta\times C\times RTT-2Sx(1-\theta)}{2-\theta}\le b\right\}\\&=P\left\{\frac{\theta\times C\times RTT-b(2-\theta)}{2S(1-\theta)}\le x\right\}\\&=\int_{\frac{\theta\times C\times RTT-b(2-\theta)}{2S(1-\theta)}}^{\infty}f_X(x)\mathrm{d}x.\end{aligned}$$

当 $w'<x\le MN$ 时,B 为变量 x 的单调函数,因此有

$$f_B(b)=\frac{2-\theta}{2S(1-\theta)}f_X\left(\frac{\theta\times C\times RTT-b(2-\theta)}{2S(1-\theta)}\right) \tag{11}$$

结合式(5)和式(11)可得:

$$f_B(b)=\frac{2-\theta}{2S(1-\theta)}\frac{1}{\sqrt{2\pi N}}\mathrm{e}^{-\frac{(\theta\times C\times RTT-2b+\theta b-2SN+2SN\theta)^2}{8S^2N(1-\theta)^2}},$$

其中,b 满足

$$\frac{\theta\times C\times RTT-2SMN(1-\theta)}{2-\theta}\le b\le\frac{\theta\times C\times RTT-2Sw'(1-\theta)}{2-\theta}.$$

为维持目标链路 100%链路利用率,我们选择最大的 b 值作为最小的缓存需求值,并且以 B_{FCCM} 表示.

$$B_{FCCM}=\frac{\theta\times C\times RTT-2Sw'(1-\theta)}{2-\theta}.$$

下面,我们将 B_{FCCM} 与 BSCL 方案中的最小缓存需求 B_{BSCL} 相比较,为分析方便,B_{BSCL} 重写如下[11]:

$$B_{BSCL}=\frac{q(N_b)\cdot C\cdot RTT-2SN_b(1-q(N_b))}{2-q(N_b)},$$

其中,

$$q(N_b)=1-\left(1-\frac{1}{N_b}\right)^{\bar{L}_{N_b}}.$$

式(4)中,N 对应于 BSCL 方案中的 N_b,因此可得 $N_b\le w'$,从而有下式成立:

$$N_b<x\le MN.$$

更进一步地,我们可得:

$$0\le\theta<q(N_b)\le 1.$$

再推理可得:

$$\frac{\theta\times C\times RTT-2Sw'(1-\theta)}{2-\theta}<\frac{q(N_b)\cdot C\cdot RTT-2SN_b(1-q(N_b))}{2-q(N_b)},$$

即 $B_{FCCM}<B_{BSCL}$ 成立,定理 3 得证. □

在 BSCL 方案中,链路的拥塞是由链路上 N 个固定的 TCP 流不断增大发送窗口,从而增大了链路上传送的数据量所导致,该拥塞区别于本文所述的流竞争拥塞.定理 3 表明,在维持链路 100%利用率的目标下,应对流竞争拥塞所需的最小缓存将不大于 BSCL 方案中的最小缓存需求.因此,在设计路由器的缓存时,BSCL 方案的最小缓存需求值可满足流竞争拥塞的缓存需求.

4 结束语

路由器需要设置缓存来应对拥塞的发生,路由器的缓存需求分析目前已经成为网络研究领域的一个热点,产生了许多具有重要影响力的研究成果,其中具有代表性的斯坦佛模型和BSCL方案都假定拥塞链路上存在固定的 N 个长 TCP 流,二者所涉及的拥塞都是由于 N 个长 TCP 流的发送窗口增长导致链路上聚合后的流量超过了链路处理能力引起的.实际上,在具体的路由器输出链路上,除了上述的拥塞以外,常见的另外一种拥塞是网络突发业务条件下由于相同目的输出链路的 TCP 流争用相同的目标链路而引发的,本文将该类型拥塞定义为流竞争拥塞.

本文的主要贡献在于:首先给出了一个新的适用于流竞争拥塞的分析模型 FCCM,借助 FCCM,我们分析了目标链路上 TCP 竞争流的分布,给出了其分布函数.分析表明,当输入链路上的 TCP 流数目 N 和路由器的输入端口数目 M 都足够大时,若 $N/M\le 5$,则目标链路上聚合后的 TCP 竞争流的流数目服从泊松分布;若 $N/M>5$,则服从正态分布.随后,本文证明了当路由器的输入链路数目 M 足够大时,目标链路上由 M 个输入链路聚合后的 TCP 竞争流的流数目分布与 M 无关这一结论.结合前面的结论,本文最后部分分析并给出了流竞争拥塞发生的条件,证明了在维持链路 100%利用率目标下,应对流竞争拥塞所需的最小缓存将不大于 BSCL 方案中的最小缓存需求这一结论,从而为路由器的缓存设置提供了理论指导.

本文中针对流竞争拥塞的缓存分析是以维持链路100%利用率为目标进行的.在下一步工作中,我们将以丢

包率和时延为分析目标,探讨路由器的缓存需求,对路由器的缓存设置给出更科学、更全面的理论指导.

References:

[1] Beheshti N, Ganjali Y, Rajaduray R, Blumenthal D, McKeown N. Buffer sizing in all-optical packet switches. In: Proc. of the OFC/NFOEC. 2006. 3–6. http://yuba.stanford.edu/buffersizing/

[2] Appenzeller G, McKeown N, Sommers J, Barford P. Recent results on sizing router buffers. In: Proc. of the Network Systems Design Conf. 2004. 18–20. http://tiny-tera.stanford.edu/~nickm/papers/index.html.

[3] Gorinsky S, Kantawala A, Turner JS. Link buffer sizing: A new look at the old problem. In: Proc. of the ISCC 2005. Washington: IEEE Computer Society, 2005. 507–514.

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

[5] 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/

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

[7] Appenzeller G, Keslassy I, McKeown N. Sizing router buffers. ACM/SIGCOMM Computer Communication Review, 2005,34(4): 281–292.

[8] Raina G, Towsley D, Wischik D. Part II: Control theory for buffer sizing. ACM/SIGCOMM Computer Communication Review, 2005,35(3):79–82.

[9] Enachescu M, Ganjali Y, Goel A, McKeown N, Roughgarden T. Part III: Routers with very small buffers. ACM/SIGCOMM Computer Communication Review, 2005,35(3):83–90.

[10] Wischik D, McKeown N. Part I: Buffer sizes for core routers. ACM/SIGCOMM Computer Communication Review, 2005,35(3): 75–78.

[11] Dhamdhere A, Jiang H, Dovrolis C. Buffer sizing for congested internet links. In: Proc. of the IEEE INFOCOM 2005. 2005. 1072–1083.

[12] Karol M, Hluchyj M, Morgan S. Input versus output queueing on a space-division switch. IEEE Trans. on Communications, 1999, 17(6):1030–1039.

[13] Devore JL. Probability and Statistics for Engineering and the Sciences. 6th ed., San Luis Obispo: BrooksCole, 2004.

[14] Chiu D, Jain R. Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. Computer Networks and ISDN Systems, 1989,17(1):1–14.

[15] Xu K, Xiong YQ, Wu JP. Analysis of Broadband IP Router Architecture. Journal of Software, 2000,11(2):179–186 (in Chinese with English abstract).

附中文参考文献:

[15] 徐恪,熊勇强,吴建平.宽带 IP 路由器的体系结构分析.软件学报, 2000,11(2):179–186.

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

兰巨龙(1962－),男,博士,教授,博士生导师,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,助教,主要研究领域为宽带信息网络,流媒体技术.

汪斌强(1973－),男,博士生,助教,主要研究领域为宽带信息网络,高速路由器核心技术.