考虑慢启动影响的TCP 吞吐量模型

韩 涛, 朱耀庭, 朱光喜, 姚文冰
(华中科技大学电信系, 湖北武汉 430074)

摘 要: TCP 吞吐量模型的研究是网络协议研究的一个重要方面, 同一些其它模型相比, Padhye 提出的TCP 吞吐量模型比较精确地描述了TCP 吞吐量与往返时间、丢包率和超时时限的关系, 但在丢包率很高的场合, Padhye 模型误差较大, 本文分析了高网络负荷下 TCP 传输的性能, 提出了一个高网络负荷下 TCP 吞吐量的改进 Padhye 模型, 实验表明, 该模型在高网络负荷环境下更接近实际情况.

关键词: TCP 吞吐量模型; 拥塞控制; 网络协议
中图分类号: TP393 文献标识码: A 文章编号: 0372-2112 (2002) 10-1481-04

A TCP Throughput Model Considering Slow Start

HAN Tao, ZHU Yao-ting, ZHU Guang-xi, YAO Wen-bing
(*Department of Electron and Information Engineer Huazhong University of Science and Technology, Wuhan, Hubei* 430074, *China*)

Abstract: Research on TCP throughput model is an important aspect of research on network protocols. Compared with some other models, the model given by Padhye is more accurate to describe the relation between TCP throughput and round trip time, rate of packet loss, and time out limit. But under heavy network load, Padhye model doesn't match the results very well. In this paper, the performance of TCP under heavy network load is analyzed, and a Padhye Model based TCP throughput model in heavy network load is given forth. The experiment indicates that this model matches the results better than Padhye model in heavy network load.

Key words: TCP throughput model; congestion control; network protocol

1 引言

今天 Internet 上的大部分流量, 包括 WWW(HTTP), 文件传送(FTP), 电子邮件(SMTP) 以及远程访问(Telnet) 都是由 TCP 协议传送的. 针对TCP 性能的研究一直是 Internet 网络协议研究的重要方面.

近两年, 已经有一些 TCP 吞吐量模型被建立起来, 其中 Padhye 模型[1] 是当前最为众所周知的, 根据本文作者的实验, Padhye 模型在网络负荷较高的场合下误差较大, 本文提出了一个在高网络负荷下TCP 吞吐量模型, 实验表明, 该模型在高网络负荷环境下更接近实际情况.

2 Padhye 的 TCP 吞吐量模型简介

无限源TCP 吞吐量主要受限于 TCP 的拥塞控制, 从 1988 年 Jacobson 提出 TCP 拥塞控制算法[2] 以来至今, TCP 拥塞控制算法已经有了很大提高, 标准的 TCP 拥塞控制(TCP/Tahoe) 包括慢启动、拥塞避免、快速重传以及重传超时的二进制回退等算法, TCP/Reno 在 TCP/Tahoe 的基础上增加了快速恢复算法[4], 目前 TCP 的实现主要是 TCP/Reno, 新的 TCP 流控和拥塞控制算法如TCP/SACK 以及 TCP/FACK 则没有普遍实现.

影响TCP 吞吐量的另一个重要因素是网络中路由器的实现, 目前 Internet 上大部分的路由器采用 Drop-tail 算法, 即, 当路由器缓冲区溢出时丢弃队列尾的数据包, 这种算法被证明是缺乏公平性和相对低效的, 一种更新的算法 RED(Random Early Detect) 将取代它.

Padhye 针对 TCP/Reno 和 Drop-tail 路由器提出 TCP 发送率及吞吐量模型, 其主要思想是将 TCP 的稳态传送过程分为 TD(Triple-duplicate, 三重重复应答) 和 TO(Time-Out, 应答超时) 两部分, 然后分别计算两部分过程的持续时间和发送的数据包, 按照如下公式计算发送率:

$$\text{发送率}=\frac{\text{TD 期间发送的数据包数}+\text{TO 期间发送的包数}}{\text{TD 持续时间}+\text{TO 持续时间}}$$

建立了 TCP 发送率 $B(p)$ 和丢失指示率 p、往返时间 RTT、累积应答因子 b、重发超时 T_0 以及最大拥塞窗口 W_{max} 之间的关系如下:

$$B(p)\approx\min\left(\frac{W_{max}}{RTT},\ \frac{1}{RTT\sqrt{\frac{2bp}{3}}+T_0\min\left(1,3\sqrt{\frac{3bp}{8}}\right)p(1+32p^2)}\right) \tag{1}$$

仿真和实验表明, Padhye 模型在较低或中等丢失指示率的情况下(网络负荷较轻的情况) 与实际情况相当符合, 比 Mahdavi 和 Floyd 的只考虑 TD 情况的简单模型[5] 精确许多, 但

收稿日期: 2001-03-09; 修回日期: 2001-11-08
基金项目: 国家"九五"重点攻关项目(No. 96-743-01-06-03)

在网络负荷较高的情况,即丢失指示率较高的情况下,Padhye 模型与实际结果仍有较大出入.

3 在高网络负荷下的 TCP 吞吐量模型

经分析,Padhye 模型对高丢失指示率情况下误差的主要原因是 Padhye 模型忽略了 TCP 拥塞算法中的慢启动阶段,而当拥塞门限大于 2 时,从 TO 阶段到 TD 阶段必然要经历一次慢启动阶段,在网络负荷较高的情况下,丢失指示率也相对较高,超时事件发生率很大,因而慢启动阶段的影响不能忽略.本文考虑到慢启动阶段的影响,提出了如下的 TCP 吞吐量模型,试验证明,该模型在丢失指示率较高的情况下比 Padhye 模型更接近实际情况.

首先,假设最大拥塞窗口 W_{max} 足够大,稳态发送率 B 只与丢失指示率 p、往返时间 RTT、累积应答因子 b、重发超时 T_0 有关.在稳态下,TCP 传输可以看作 TD、TO 和 SS(Slow Start,慢启动)三种过程的交替,如图 1 所示.图中纵轴 W 为拥塞窗口尺寸,横轴 t 为时间.

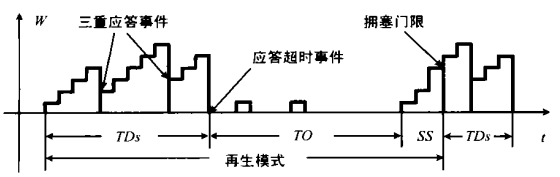

图 1 TCP 传输模式

这里 TD 过程是一个或多个连续的拥塞避免 TDi 过程构成的,而 TO 过程后面一定是 SS 过程,所以这样的再生模式构成了具有 TDi 和 TO-SS 两种状态的 Markov 链,如图 2 所示.状态转移矩阵为:

$$P=\begin{bmatrix} 1-p_{TD\to TOSS} & p_{TD\to TOSS} \\ p_{TOSS\to TD} & 1-p_{TOSS\to TD} \end{bmatrix} \tag{2}$$

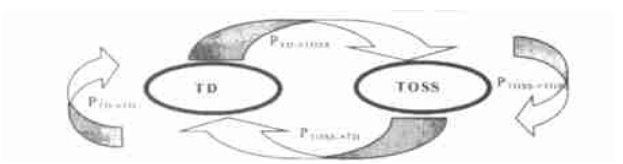

图 2 TCP 再生模式状态变迁图

这里 $p_{TD\to TOSS}$ 是从 TD_i 过程到 TO-SS 过程的转移概率,$p_{TOSS\to TD}$ 是从 TO-SS 过程到 TD_i 过程的转移概率,根据 Padhye 模型,有:

$$E[W]=\frac{2+b}{3b}+\sqrt{\frac{8(1-p)}{3bp}+\left(\frac{2+b}{3b}\right)^2}=\sqrt{\frac{8}{3bp}}+o(1/\sqrt{p})\approx\sqrt{\frac{8}{3bp}} \tag{3}$$

这里 W 是表示在 TD 过程中拥塞窗口大小的随机变量,由 Padhye 的推导,有:

$$p_{TD\to TOSS}=\sum_{w=1}^{\infty}\min\left(1,\frac{(1-(1-p)^3)(1+(1-p)^3(1-(1-p)^{w-3}))}{1-(1-p)^w}\right)P[W=w]$$
$$\approx\min\left(1,\frac{(1-(1-p)^3)(1+(1-p)^3(1-(1-p)^{E[W]-3}))}{1-(1-p)^{E[W]}}\right)$$
$$\approx\min\left(1,\frac{3}{E[W]}\right)=\min\left(1,3\sqrt{\frac{3bp}{8}}\right)$$

因此有下式成立:

$$P_{TD\to TD}=1-P_{TD\to TOSS}\approx\max\left(0,1-3\sqrt{\frac{3bp}{8}}\right)$$

在 SS 过程中,拥塞窗大小是以指数形式增长的,当窗口增大到拥塞门限(即前一个 TD 过程中最后一个来回中拥塞窗口的一半)时,将直接进入 TD 过程,设其概率为 P',有:

$$P'=\sum_{w=1}^{\infty}\left(\prod_{i=1}^{w/2}(1-p)^{bi^2}\right)P[W=w]\approx\sum_{i=1}^{\sqrt{E[W]/2}}(1-p)^{bi^2}$$
$$=(1-p)^{\frac{1}{24}b(\sqrt{2E[W]}+2)(\sqrt{2E[W]}+1)\sqrt{2E[W]}}\approx(1-p)^{\frac{b}{6}\sqrt{\frac{8}{3bp}}} \tag{4}$$

在窗口增大到拥塞门限之前,可能因为包丢失而进入 TO 或 TD 状态,设进入 TO 状态的概率为 P'',有:

$$p''=\sum_{w=1}^{\infty}\min\left(1,\frac{(1-(1-p)^3)(1+(1-p)^3(1-(1-p)^{w-3}))}{1-(1-p)^w}\right)P[W_{ss}=w]$$
$$\approx\min\left(1,\frac{(1-(1-p)^3)(1+(1-p)^3(1-(1-p)^{E[W_{ss}]-3}))}{1-(1-p)^{E[W_{ss}]}}\right)$$
$$\approx\min\left(1,\frac{3}{E[W_{ss}]}\right) \tag{5}$$

这里 E_{ss} 为 SS 过程中拥塞窗口大小期望值,显然:

$$E[W_{ss}]=\sum_{i=1}^{\log_2\frac{E[W]}{2}}2^iP[W_{ss}=w]\approx\frac{1}{\log_2\frac{E[W]}{2}}\sum_{i=1}^{\log_2\frac{E[W]}{2}}2^i$$
$$=\frac{E[W]-2}{\log_2\frac{E[W]}{2}}$$

考虑到拥塞窗口最小为 1,所以:

$$E[W_{ss}]=\max\left(1,\frac{E[W]-2}{\log_2\frac{E[W]}{2}}\right) \tag{6}$$

由式(4)、(5)和(6)可得:

$$P_{TOSS\to TD}=P'+(1-P')(1-P'')\approx(1-p)^{\frac{b}{6}\sqrt{\frac{8}{3bp}}}$$
$$+(1-(1-p)^{\frac{b}{6}\sqrt{\frac{8}{3bp}}})(1-\min\left(1,\frac{3}{E[W_{zz}]}\right)$$
$$\approx(1-p)^{\frac{b}{6}\sqrt{\frac{8}{3bp}}}$$

以及:

$$P_{TOSS\to TOSS}=1-P_{TOSS\to TD}\approx1-(1-p)^{\frac{b}{6}\sqrt{\frac{8}{3bp}}}$$

所以其稳态分布为:

$$P_{TD}=\frac{P_{TOSS\to TD}}{P_{TD\to TOSS}+P_{TOSS\to TD}}=\frac{(1-p)^{\frac{b}{6}\sqrt{\frac{8}{3bp}}}}{\min\left(1,3\sqrt{\frac{3bp}{8}}\right)+(1-p)^{\frac{b}{6}\sqrt{\frac{8}{3bp}}}} \tag{7}$$

$$P_{TOSS}=\frac{P_{TD\to TOSS}}{P_{TD\to TOSS}+P_{TOSS\to TD}}=\frac{\min\left(1,3\sqrt{\frac{3bp}{8}}\right)}{\min\left(1,3\sqrt{\frac{3bp}{8}}\right)+(1-p)^{\frac{b}{6}\sqrt{\frac{8}{3bp}}}} \tag{8}$$

下面求 SS 过程持续时间的数学期望.设 SS 过程持续时间为随机变量 A_{ss},来回个数为随机变量 X_{ss},因为 SS 过程中拥塞窗大小按指数增长的,对于 SS 过程的最后一个来回,拥塞窗口达到拥塞门限(即 TD 过程中拥塞窗口的一半),有:

$$E[X_{ss}]\approx\max\left(1,b\cdot\log_2\frac{E[W]}{2}\right) \tag{9}$$

以及

$$E[A_{ss}]\approx RTT\cdot E[X_{xx}]=RTT\max\left(1,b\cdot\log_2\frac{E[W]}{2}\right) \tag{10}$$

设 Y_{ss} 为一个 SS 过程中发送的总数据包个数的随机变量, 有:

$$E[Y_{ss}] \approx E[W_{ss}]E[X_{ss}] \approx \max(1, b(E[W]-2)) \quad (11)$$

最后, 考虑了 SS 过程的 TCP 流量模型的发送率如下:

$$\text{发送率} = \frac{\text{TD 期间发送的数据包数} + \text{TOSS 期间发送的数据包数}}{\text{TD 持续时间} + \text{TOSS 持续时间}}$$

按照 Padhye 模型有: 单个 TD 过程中发送的数据包个数的数学期望 $E[Y] = \frac{1-p}{p} + E[W]$, 单个TD 过程包含的来回数的数学期望为 $E[X] \approx \sqrt{\frac{2b}{3p}}$, 单个 TD 过程持续时间的数学期望为 $E[A] \approx RTT\sqrt{\frac{2b}{3p}}$, 单个 TO 过程持续时间的数学期望 $E[A_{TO}] = T_0\frac{1+p+2p^2+4p^3+8p^4+16p^5+32p^6}{1-p}$, 单个 TO 过程发送的数据包个数的数学期望 $E[R_{TO}] = \frac{1}{1-p}$.

由上述等式以及式(10)和式(11), 发送率 $b(p)$ 为:

$$b(p) = \frac{P_{TD}E[Y] + P_{TOSS}(E[R_{TO}] + E[Y_{ss}])}{P_{TD}E[A] + P_{TOSS}(E[A_{TO}] + E[A_{SS}])} = \frac{P_{TOSS\to TD}E[Y] + P_{TD\to TOSS}(E[R_{TO}] + E[Y_{ss}])}{P_{TOSS\to TD}E[A] + P_{TD\to TOSS}(E[A_{TO}] + E[A_{SS}])}$$

$$= \frac{(1-p)^{\frac{b}{6\sqrt{bp}}}\left(\frac{1-p}{p} + \sqrt{\frac{8}{3bp}}\right) + \min\left(1, 3\sqrt{\frac{3bp}{8}}\right)\left(\frac{1}{1-p} + \max\left(1, 2b\left(\sqrt{\frac{2}{3bp}} - 1\right)\right)\right)}{(1-p)^{\frac{b}{6\sqrt{bp}}}RTT\sqrt{\frac{2b}{3p}} + \min\left(1, 3\sqrt{\frac{3bp}{8}}\right)\left(T_0\frac{f(p)}{1-p} + RTT\max\left(1, b\log_2\sqrt{\frac{2}{3bp}}\right)\right)}$$

这里

$$f(p) = 1 + p + 2p^2 + 4p^3 + 8p^4 + 16p^5 + 32p^6.$$

考虑最大拥塞窗口限制, 有:

$$b(p) = \min\left(\frac{W_{\max}}{RTT}, G(p)\right) \quad (12)$$

其中:

$$G(p) = \frac{(1-p)^{\frac{b}{6\sqrt{bp}}}\left(\frac{1-p}{p} + \sqrt{\frac{8}{3bp}}\right) + \min\left(1, 3\sqrt{\frac{3bp}{8}}\right)\left(\frac{1}{1-p} + \max\left(1, 2b\left(\sqrt{\frac{2}{3bp}} - 1\right)\right)\right)}{(1-p)^{\frac{b}{6\sqrt{bp}}}RTT\sqrt{\frac{2b}{3p}} + \min\left(1, 3\sqrt{\frac{3bp}{8}}\right)\left(T_0\frac{f(p)}{1-p} + RTT\max\left(1, b\log_2\sqrt{\frac{2}{3bp}}\right)\right)}$$

这里, 将考虑了 SS 过程影响的 TCP 流量模型称为 SS-Padhye 模型, 式(12)即为该模型的流量公式.

4　试验结果及分析

在一系列使用无限源的 TCP 连接中, 测试结果表明, SS-Padhye 模型在丢失指示率 p 较小或中等时与 Padhye 模型差别很小, 在丢失指示率为较大和很大时, 比 Padhye 模型更接近实际情况.

实验环境为中国教育与科研网(CERNET), 实验涉及到的主机分布于武汉、北京、广州、西安、南京、上海、合肥、沈阳以及北美, , 主机使用的操作系统包括 Windows 9x、Windows NT、Linux、FreeBSD 以及 Solaris 等, 主机采用的联网方式涉及以太网、电话拨号、ISDN 以及 Cable Modem 等, 实验的时间选择为从 1998 年到 2000 之间的一些随机时间, 包括白天和夜晚、工作日和公休日. 实验手段主要采用从源主机建立到目标主机的 discard 口的 TCP 连接或者通过编写测试程序完成, 在源主机所在的本地网络采用 tcpdump 捕获网络上传送的 TCP 数据包及应答包, 通过对捕获的包进行分析, 得到了数十组实验结果.

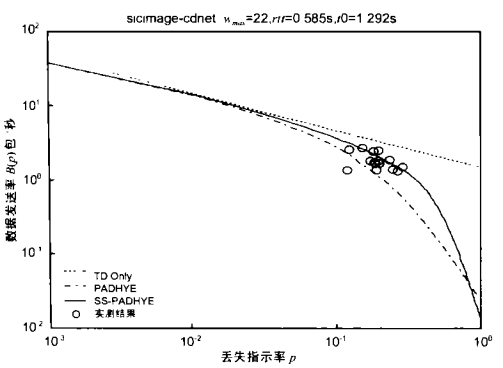

图 3　测试结果(一)

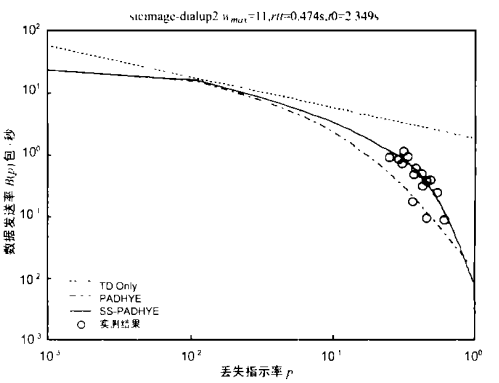

图 4　测试结果(二)

图 3 和图 4 是其中两个在较高丢失率下的典型实验结果, 图中虚线表示只考虑了 TD 过程的 TCP 流量模型(TD Only)计算出来的TCP 发送率, 点划线表示按 Padhye 模型计算的发送率, 实线表示按 SS-Padhye 模型计算的发送率, 小圆圈表示实际测量的结果, 每个小圆圈代表在一段持续时间为 100s 的 TCP 连接中平均数据发送率和丢失指示率之间的关系.

从图中可以看出, 在丢失指示率较小的情况下, Padhye 模型和 SS-Padhye 模型差别很小; 在丢失指示率较大的情况下, SS-Padhye 模型计算出来的 TCP 发送率大于 Padhye 模型计算出来的 TCP 发送率, 其原因是在丢失指示率较大的情况下, TO 过程的发生率也较大, 而 Padhye 模型忽略了跟随在每一个 TO 过程 SS 过程, 因而估算的发送率偏小; 在丢失指示率很大的情况下, 由 SS 过程直接转向 TO 过程的概率也大大增加, 因而 TCP 发送率也相应大大减小, 此时考虑了 SS 过程的 SS-Padhye 模型估算的 TCP 发送率小于 Padhye 模型估算的发送

率.所以 SS-Padhye 模型在高丢失指示率的情况下比 Padhye 模型更符合实际情况,实验结果也证实了这一点.

5 结论

理论推导和实验验证两方面的工作均表明在 Padhye 模型基础上考虑了慢启动过程的 SS-Padhye 模型的 TCP 吞吐量估计值在高丢失指示率的环境下比 Padhye 模型更接近实际情况,因而本文的工作在设计高丢失指示率环境下的 TCP 友好的传输控制协议具有指导意义.

根据 SS-Padhye 模型,作者设计了一个基于率控制的 TCP 友好的多媒体传输协议,实现结果表明,该协议达到了理想的 TCP 友好性的设计目标.

参考文献:

[1] Padhye J, Firoiu V, Towsley D, Kurose J. Modeling TCP Reno Performance: A Simple Model and its Empirical Validation [A]. Proceedings of SIGCOMM'98 [C]. SIGCOMM, 1998.

[2] Jacobson V. Congestion avoidance and Control [A]. Proceedings of SIGCOMM'88 [C]. SIGCOMM, 1988. 314-329

[3] 潘建平,顾冠群,吴国新. TCP/IP 拥挤控制研究 [J]. 小型微型计算机系统, 1999, (20) 4: 251-255.

[4] Stevens W. TCP Slow Start, Congestion Avoidance, Fast retransmit, and Fast Recovery Algorithms [S]. RFC2001, DDN Network Information Center, Jan 1997.

[5] Mahdavi J and Floyd S. TCP-friendly unicast rate-based flow control [Z]. Note sent to end2end-interest mailing list, 1997.

[6] Bolliger J, Gross T, Hengartner U. Bandwith modeling for network-aware applications [A]. Proceedings of INFOCOMM'99 [C]. INFOCOMM, 1999

作者简介:

韩 涛 男,1970 年生于湖北天门,讲师,现为华中科技大学电信系博士研究生,研究方向为多媒体信息处理、计算机网络通信.

朱耀庭 男,1939 年生于湖北长沙,教授,博士生导师,1961 年毕业于华中工学院无线电工程系.

朱光喜 男,1945 年生于广西玉林,教授,博士生导师,1969 年毕业于华中工学院无线电工程系,曾主持完成国家自然科学基金、部委基金和"九五"国家重点攻关项目十余项,已发表学术论文 100 余篇.