软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(10):2346−2361 [doi: 10.13328/j.cnki.jos.004461]

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

基于抽样流记录的 RTT 估计*

苏 琪[1,2], 龚 俭[1,2], 苏艳珺[1,2]

[1](东南大学 计算机科学与工程学院,江苏 南京 210096)

[2](江苏省计算机网络重点实验室,江苏 南京 210096)

通讯作者: 苏琪, E-mail: qsu@njnet.edu.cn

摘 要: 往返时延(RTT)是网络测量中的一个重要测度,是刻画网络性能的重要指标.传统的 RTT 测量都是基于报文的,需要专门的主动或被动测量平台的支持.提出一种新的 RTT 估计方法,仅使用现有路由器设备提供的流记录,不需要额外的网络测量设施.通过对 TCP 块状流传输特性的分析,分别建立了当套接字缓冲区长度与带宽延迟积 BDP 相对较小、较大和相近这 3 种情况下的 RTT 估计模型.实验结果表明,这些模型都能很好地完成 RTT 估计.同时,由于在估计当中只使用了流持续时间和总报文两个变量,因此,该方法同样适用于以抽样流记录为输入的环境,能够有效地应用于现有的大规模主干网环境的网络检测与管理.

关键词: RTT 估计;抽样流记录;网络测量;网络管理;随机过程

中图法分类号: TP393

中文引用格式: 苏琪,龚俭,苏艳珺.基于抽样流记录的 RTT 估计.软件学报,2014,25(10):2346−2361. http://www.jos.org.cn/1000-9825/4461.htm

英文引用格式: Su Q, Gong J, Su YJ. RTT estimation based on sampled flow data. Ruan Jian Xue Bao/Journal of Software, 2014, 25(10):2346−2361 (in Chinese). http://www.jos.org.cn/1000-9825/4461.htm

RTT Estimation Based on Sampled Flow Data

SU Qi[1,2], GONG Jian[1,2], SU Yan-Jun[1,2]

[1](School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

[2](Jiangsu Provincial Key Laboratory of Computer Network Technology, Nanjing 210096, China)

Corresponding author: SU Qi, E-mail: qsu@njnet.edu.cn

Abstract: Round-Trip time (RTT) is an important metric for network measurement and an essential indicator for network performance monitoring. Traditional packet trace based RTT estimation usually depends on particular active or passive measurement platforms. This paper proposes a new RTT estimation method, which merely takes flow data from existed routers and hardly needs extra network measurement facility. Based on the analysis of transmission features of TCP bulk flow, RTT estimation models are established corresponding to the conditions where socket buffer size and bandwidth delay product (BDP) are relatively small, large and approximate. Experiments show RTT estimation can be well accomplished through those models. Moreover, considering only duration and total packet number of a TCP bulk flow are involved in estimation, this method is also adoptable to situation with sampling flow data as input, and thus is effective in monitoring and managing the large-scale backbone network performance.

Key words: RTT estimation; sampled flow data; network measurement; network management; stochastic process

往返时延(round-trip time,简称 RTT)是刻画网络性能的一项重要指标,传统的 RTT 可以被定义为一个报文发送时戳与收到这个报文对应的应答报文时戳的间隔.一个 RTT 的值,是对特定的两个 IP 地址之间的传输性能的重要衡量标准之一.

* 基金项目: 国家科技支撑计划(2008BAH37B04); 国家基础研究发展计划(973)(2009CB320505); 国家自然科学基金(60973123)
收稿时间: 2013-03-07; 定稿时间: 2013-07-30

对于网络管理,周期性的 RTT 测量可为网络管理员提供网络的整体性能视图.例如:一个 ISP 在评估自己的服务质量时,需要了解与其他 ISP 的时延情况;对于分布在各地的企业网,需要测量它们之间的网络延迟来评估业务的运行状况;对于一个大型的网络内容提供商(ICP),管理人员需要监测各地区访问这个网站的 RTT 以评估服务质量.

RTT 的估计是一个经典问题,已有大量的相关研究,大体上可分为基于主动测量或基于被动测量两类方法.

对于主动测量,最常见的是使用发送 ICMP 回显请求和收到 ICMP 回显应答的间隔[1],或者是发出 TCP SYN 报文与收到 TCP SYN-ACK 报文的间隔作为对 RTT 的估计.主动测量方法能达到很高的精度,但是测量只能在端系统中进行,而目前的网络管理一般的测量环境在中间的网络上.更进一步地,随着网络安全问题的日益突显,对于非协作的对端网络管理者,很可能会将这种周期性的测量流量视为恶意流量而予以过滤,使测量不能正常进行.

被动测量基于对实际流量的观测,经典的方法有 SA 算法和 SS 算法[2].SA 算法通过 TCP 请求连接时,访问端(caller)向被访问端(callee)发送的最后一个 SYN 报文与第 1 个 ACK 报文的间隔对 RTT 进行估计;而 SS 算法的基本思想是,认为第 1 轮(burst)TCP 报文与第 2 轮的间隔可以近似地等于这个 TCP 连接的 RTT,文献[3]也使用了类似的思想.Zhang 等人[4]提出一种非统计的 RTT 估计方法:他们生成一系列候选的 IP 间 RTT 值,然后用最符合当前 TCP 会话行为的候选值来估计真正的 RTT.Jaiswal 等人[5]先计算出拥塞窗口(cwnd),再根据拥塞窗口找出发送端发出的触发接收端发送 ACK 报文的特定报文,而它们之间的间隔即为 RTT 的估计值.张佚博等人[6]提出的 PRE 算法根据在同一轮次的报文间隔与轮次间的报文间隔显著不同这一特点来区分两种不同的间隔,进而得出 RTT 的估计值.上述几种被动测量方法都是基于全报文的,在大型高速网络中,采集全报文进行分析需要特殊的解决方案,这些方案需要专用设备的支持.同时,采集到的数据必须通过离线分析,不能用于网络性能实时监控.

如今,大多数主流的路由器均提供了采集流记录的机制,如 NetFlow[7],sFlow[8]等,也出现了相应的互联网标准 IPFIX[9].Strohmeier 等人[10]提出基于流记录的 RTT 测量模型,利用符合一定条件的两条方向相反的流记录的起始时间的间隔作为对 RTT 的估计,这种方法不允许使用抽样的流记录,而且需要在流记录之间进行流关联,因此缺乏可扩展性和实用性.从精细化网络管理的需要出发,基于实际主干网环境中路由器设备提供的抽样流记录,本文提出了一种 RTT 测量的新方法.之前有大量的工作对于存在丢包的 TCP 传输行为进行建模[11,12],本文借鉴它们的思想,对于每一条符合条件的单向 TCP 流记录,得出相应的关于 RTT 的计算模型,从而得到对这条流整个传输过程中的 RTT 平均值的估计.

图 1 所示为一个 TCP 流的会话过程.本文把 TCP 传输抽象成 3 部分:发送端、接收端和中间的网络,而流记录的采集发生在中间的网络中的观测点.在 TCP 传输期间,发送端发送报文的密度是非均匀的,而接收端发送报文的密度受其影响,也是非均匀的.在有充足数据要发送的情况下,发送端连续发送一簇(burst)n_1 个报文到接收端,并达到拥塞窗口的最大值后不再发送报文;报文到达后,接收端经过 t_1 时间的处理,发送 ACK 报文给发送端确认;发送端经过 t_2 时间的处理,发送一簇 n_2 个报文,同时发送端再一次达到拥塞窗口的最大值.以此类推,直至传输完成.如图 1 中虚线框所示,本文称 TCP 每次发送一簇报文并收到相应的 ACK 的过程为一个轮次.

本文中 RTT 估计的数据来源于在中间路由器(图 1 中的观测点位置)上测量到的 TCP(抽样的)流记录,因此,测量只能是对相邻两次发送连续报文的间隔(如图 1 中的 RTT_2)的平均值进行估计,而这与真实的 RTT(如图 1 中的 RTT_1)存在一定的差距,这主要是发送端和接收端的处理时间(图 1 中的 t_1,t_2)的影响所致.为了提高估计精度,必须使这些时间尽可能地短.

TCP 的流主要分成两类,那些包含块状数据(如 FTP 协议、HTTP 协议的数据)的流称为 TCP 块状流,而包含交互数据(如 Telnet 和 Rlogin 的数据)的流称为 TCP 交互流[1].TCP 交互流中,t_1 通常包含系统处理命令(如 Telnet),t_2 通常包含等待用户的键入等时间,这些时间相对于 RTT 的值不能被忽略;而对于 TCP 块状流,一旦发送端准备好数据,之后的传输过程一般没有数据处理和人为干预的时间,能很好地避免 t_1,t_2 的影响.因此,本文选取 TCP 块状流作为估计 RTT 的数据源.

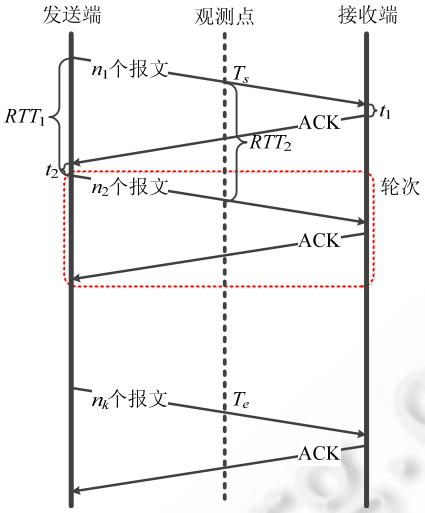


Fig.1 Basic model for RTT estimation

图 1 RTT 估计的基本模型

本文估计 RTT 的基本思想是:TCP 块状流传输时,在有数据的情况下,均是以当前的拥塞窗口作为簇长来传输的,因此,图 1 中的 $n_1,n_2,\ldots,n_k$ 即对应了每个轮次中的 *cwnd* 大小.令 $w(r)$是关于轮次 r 的表示 *cwnd* 大小的函数,那么,当传输的总报文数为 N、传输持续时间为 T 时,传输过程中的平均 RTT 为 τ,总轮次数为 T/τ,则有:

$$\sum_{r=1}^{T/\tau} w(r) = N \tag{1}$$

因此,如果可以估计出函数 $w(r)$,即可通过解方程(1)来求出平均 RTT 的估计值 τ.

本文通过分析 TCP 块状流的传输特性,分别建立了以下 3 种情况下的 RTT 估计模型:

1) 当套接字缓冲区大小相对于带宽延迟积 BDP(bandwidth-delay product)较小时;
2) 当套接字缓冲区大小相对于 BDP 较大时;
3) 当套接字缓冲区大小与 BDP 相近时.

实验结果表明,在 3 种不同的情况下,各自的模型都能很好地完成 RTT 估计.同时,由于在估计当中只使用了流持续时间和总报文两个变量,因此,本方法同样适用于以抽样流记录为输入的环境,能够有效地应用于大规模高速网络的检测与管理.

本文第 1 节列出关于 TCP 的几点主要假设.第 2 节讨论 TCP 块状流的传输行为.在此基础上,第 3 节分 3 种情况分别建立估计 RTT 的模型,同时讨论抽样对于估计的影响.第 4 节通过仿真和实测对各种估计模型进行评估.第 5 节总结本文的工作.

1 关于 TCP 的主要假设

1.1 关于TCP块状流传输过程的假设

基于上述讨论,本文假设在 TCP 块状流传输过程中,发送端和接收端的处理时间为 0.同时,由于 TCP 块状流的传输机理,簇内的相邻报文的时间差相较于簇间的时间差很小,因此假设同一簇的报文是同时从发送端发出的,且同时到达观测点,并假设每一簇报文从发送端到观测点(流采集点)所耗费的时间基本相同,因此,可以认为在观测点上观测到的簇间的时间差即为 RTT.

1.2 忽略TCP接收端影响

在网络的协议栈中有两个重要的缓冲区:一个是在内核用于存储数据的“套接字缓冲区(socket buffer)”,一

个是应用层用于存储数据的“应用层缓冲区(application buffer)”.当应用程序需要发送数据时,通过调用 send(·)函数将应用层缓冲区的数据复制到套接字缓冲区中,再由内核依据拥塞控制策略传输数据.当套接字缓冲区的数据全都发送完后,再从应用层缓冲区继续复制数据至套接字缓冲区,继续发送数据.我们假设发送端的套接字缓冲区长度始终不大于接收端的套接字缓冲区长度.这是一个合理的假设,因为在多数系统中,发送端的套接字缓冲区的默认长度都小于接收端缓冲区(如在 Linux 2.6.27 中,发送端的套接字缓冲区的长度为 16K 字节,而接收端的套接字缓冲区长度为 85K 字节).因此在之后的讨论中,忽略接收端的影响,只考虑发送端.

1.3 流传输过程中丢包率恒定假设

网络中,丢包主要包括链路层丢包和路由器队列丢包两种,由于现在的网络设施可靠性有所增加,链路层丢包出现的概率已经很低,可以忽略不计,因此,本文主要讨论的是传输经过的路由器中队列的丢包,这种丢包主要是由于传输的数据压力超过了传输路径中的某个路由器的处理能力造成的.宏观来说,这种丢包情况与传输压力存在一定的相关性,传输压力大时,丢包情况增多,相应的丢包率增大;反之,丢包率减小.正常情况下,网络中的数据传输压力在小时间粒度上,例如分钟级,是基本稳定的.因此,可以假设在这样的时间粒度上丢包率也是基本稳定的.进一步地,假设在一条流的传输时间内,丢包是一个强度(intensity)恒定的泊松过程(Poisson process),这个假设在文献[13]中同样用到,此时,丢包发生的次数即假设为这个过程在传输时间的期望值.

1.4 流传输过程中RTT恒定假设

RTT 在流传输中具有波动性,但在宏观的网络性能监测中,观测这种波动意义很有限,与测量这种波动性所需花费的代价不成比例.事实上,由于流传输的时间相对较短,在这么短的时间内,网络性能反映在 RTT 上的差异较小,波动性的主要成因是端系统的处理时间,通过上一节的分析,用于本文测量的流记录的端系统处理时间很短,因此,RTT 在流传输中的波动性不大.

综上,在测量中,本文关注的是在流传输过程中的 RTT 的平均值.为了建模方便,本文假设在一条流的传输过程中 RTT 是恒定的,而这个 RTT 就是本文需要估计的测度.

2 丢包和带宽延迟积 BDP 对 TCP 块状流传输的影响

由于假设中忽略了接受端的影响,因此本文主要对发送端进行讨论.令发送端的套接字缓冲区大小为 B 字节,则对应的报文数 W 为

$$W=\frac{B}{MSS} \tag{2}$$

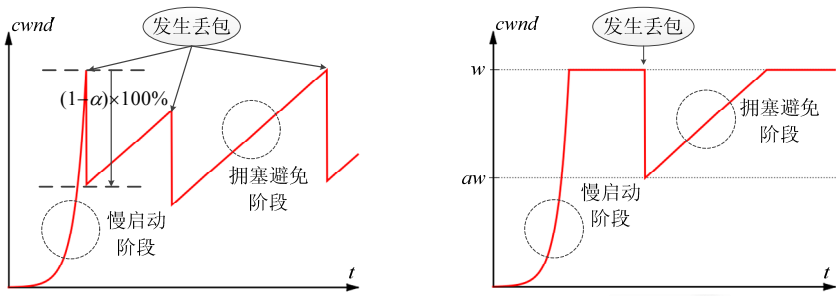
其中,MSS 为最大段长(maximum segment size).

图 2 所示的是 TCP 传输过程中,两种情况下发生丢包时拥塞窗口变化的典型曲线.拥塞窗口长度(方便起见,本文中的拥塞窗口大小用报文个数来描述)反映的是没有收到任何应答情况下能传输的最大的报文个数.传输开始时,首先是慢启动阶段,拥塞窗口呈指数增长.当发生丢包时,拥塞窗口下降一定的比例,同时进入拥塞避免阶段,并按照 AIMD 的策略,若无丢包,则待当前的报文全部被确认后,窗口长度增加一个常数;若有丢包发生,则窗口立即下降一定的比例.但是 $cwnd$ 的增长会受到 W 的限制,当 $cwnd$ 到达了 W 后,如果不发生丢包,则会保持在 W,直到发送丢包降低一定比例为止(如图 2(b)所示).

实际的 TCP 传输行为与带宽延迟积 BDP 有关.当套接字缓冲区大小(W)相对于 BDP 较大时,可以认为 TCP 传输中拥塞窗口总是小于套接字缓冲区大小,因此几乎不会受到套接字缓冲区大小的限制,即,传输行为如图 2(a)所示.

而当套接字缓冲区大小相对于 BDP 较小时,拥塞窗口将会很容易达到套接字缓冲区大小(W),并保持这个大小,直到丢包发生.当出现丢包后,拥塞窗口下降一定比例.但是由于 BDP 很大,拥塞窗口又迅速地恢复到原来的大小,如图 2(b)所示.

最后,当套接字缓冲区大小与 BDP 相近时,则图 2(a)和图 2(b)所示两种情况均有可能出现.



(a) 拥塞窗口小于 W (b) 拥塞窗口等于 W

Fig.2 Typical window curve for TCP with packet loss when *cwnd* is less than W and equal to W

图 2 当拥塞窗口小于 W 和等于 W 时发生丢包,TCP 拥塞窗口变化曲线

3 RTT 的估计模型

本文分 3 种情况来讨论相应的 RTT 的估计:(1) 当套接字缓冲区大小相对于 BDP 较小时;(2) 当套接字缓冲区大小相对于 BDP 较大时;(3) 当套接字缓冲区大小与 BDP 相近时.

3.1 当W相对于BDP较小时

因 W 较小,连续发送的报文数相应较少,更不易形成突发拥塞.因此可以假设,在慢启动阶段直至 *cwnd* 增长为 W 之前不存在丢包;同时,每次丢包的恢复阶段直至 *cwnd* 增长为 W 之前不会发生新的丢包,直至 *cwnd* 恢复到 W.

由于丢包率稳定这一假设,可令一个 TCP 连接的传输过程中稳定的丢包率为 p;同时,总报文数为 N,则可以估计出丢包数 N_{loss}:

$$N_{loss}=Np \tag{3}$$

图 3 是当 W 相对于 BDP 较小时,TCP 传输过程中 *cwnd* 变换的示意图(图中,横轴的单位为轮次),图中实线以下的面积为传输过程中的总报文数,而 S_{SS} 和 S_{CA} 分别对应于慢启动区域(图 3 中 S_{SS} 所在的阴影区域)和丢包区域(图 3 中 S_{CA} 所在的阴影区域)的面积.当发生了 N_{loss} 次丢包时,则共会出现 N_{loss} 个面积为 S_{CA} 的区域.由此,以 W 为长、以传输持续时间 T 为宽的长方形总面积可以表示为

$$W\frac{T}{\tau}=N+S_{SS}+S_{loss}=N+S_{SS}+N_{loss}S_{CA} \tag{4}$$

其中,τ为传输过程中的平均 RTT.

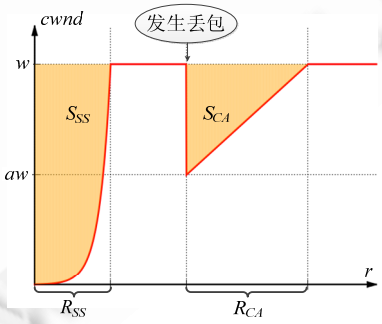


Fig.3 Sketching RTT estimation when W is small compared with BDP

图 3 当 W 相对于 BDP 较小时 RTT 估计示意

在慢启动阶段,第 r(相对于慢启动阶段的开始时的轮次)轮次的拥塞窗口大小 w_{SS} 为

$$w_{SS}(r)=2^r \tag{5}$$

又,$w_{SS}(R_{SS})$等于 W,则慢启动的总轮次 R_{SS} 为

$$R_{SS}=\log_2 W \tag{6}$$

则 S_{SS} 的计算如下:

$$S_{SS}=\sum_{r=0}^{R_{SS}}(W-w_{SS}(r))=W\log_2 W-2W+1 \tag{7}$$

在拥塞避免阶段中,每隔一个 RTT 时间,拥塞窗口增加一个常数 k(如 1[14]),则第 r(相对于当前的拥塞避免阶段的开始时的轮次)轮次的拥塞窗口 $w_{CA}(r)$为

$$w_{CA}(r)=\alpha W+kr \tag{8}$$

其中,α表示发现丢包后,拥塞窗口调整后的值与原值的比.

由于 $w_{CA}(R_{CA})$等于 W,则拥塞避免阶段经历的总轮次为

$$R_{CA}=\frac{(1-\alpha)W}{k} \tag{9}$$

则 S_{CA} 的计算如下:

$$S_{CA}=\frac{(1-\alpha)W}{2}R_{CA}=\frac{(1-\alpha)^2W^2}{2k} \tag{10}$$

最终,综合上述各式可得 RTT 的估计值τ:

$$\tau=\frac{WT}{N+W\log_2 W-2W+1+Np\dfrac{(1-\alpha)^2W^2}{2k}} \tag{11}$$

其中,W 为发送端 Socket 缓冲区大小,T 为流持续时间,N 为总报文数,p 为丢包率,α,k 为 TCP 具体拥塞控制策略中的相关常量.

3.2 当W相对于BDP较大时

当 W 相对于 BDP 较大时,传输有两个阶段:慢启动阶段直至发送第 1 次丢包以及之后的拥塞避免阶段(如图 2(a)所示).慢启动阶段的模型和上一节一致,但是拥塞避免阶段的模型却较为复杂,因为拥塞窗口将随着丢包的发生而变化,虽然丢包率是恒定的,但是丢包的发生是随机的,因此可使用随机过程进行建模.

令丢包是一个强度(intensity)为λ的泊松过程(Poisson process):$\{L(t)\}$,则在总传输时间为 T、总报文数为 N、丢包率为 p 的 TCP 传输中,假设丢包数为 $L(t=T)$的期望值,则有:

$$E(L(T))=\lambda T=Np \tag{12}$$

即:

$$\lambda=\frac{Np}{T} \tag{13}$$

定义随机过程$\{X_n\}$,表示第 n 次丢包前瞬间拥塞窗口的大小.令β为在拥塞避免阶段中,每隔一个 RTT 时间,拥塞窗口增加的常数与 RTT 的比值(k/τ),则有:

$$X_{n+1}=\alpha X_n+k\frac{S_n}{\tau}=\alpha X_n+\beta S_n \tag{14}$$

其中,S_n 为第 n 次丢包与第 n+1 次丢包的时间间隔.根据文献[15]中的定理 2.1,有 $S_n(n=1,2,3,\ldots)$是 i.i.d.的;同时,服从于参数为λ的指数分布.

文献[16]对形如:

$$Y_{n+1}=A_nY_n+B_n \tag{15}$$

的随机等式进行了深入的讨论,其中,文中定理 1 为:当$\{A_n\}$,$\{B_n\}$是平稳的(stationary)、遍历的(ergodic)随机过程,同时满足条件(其中,$(\cdot)^+$表示取 max(·,0)):

$$-\infty\leqslant E\log|A_0|<0 \text{ and } E(\log|B_0|)^+<\infty,$$

或者满足条件:

$$P(A_0=0)>0,$$

则公式(15)存在唯一稳定解(stationary solution):

$$y_n=\sum_{j=0}^{\infty}\left(\prod_{i=n-j}^{n-1}A_i\right)B_{n-j-1},$$

即下式成立:

$$P(\lim_{n\to\infty}|Y_n-y_n|=0)=1.$$

对于公式(14),$\{A_n=\alpha\}$,$\{Bn=\beta S_n\}$,假设 RTT 恒定,$\alpha(0<\alpha<1)$、β是常数,而$\{S_n\}$是 i.i.d 的,显然有$\{A_n\}$,$\{B_n\}$是平稳的、遍历的随机过程,且 $E\log|A_0|=\log\alpha<0$,$E(\log|B_0|)^+=(\log\beta\lambda)^+<\infty$,因此,存在一个稳定解 x_n:

$$x_n=\sum_{j=0}^{\infty}\left(\prod_{i=n-j}^{n-1}\alpha\right)\beta S_{n-j-1}=\beta\sum_{i=0}^{\infty}\alpha^i S_{n-i-1} \tag{16}$$

使得:

$$P(\lim_{n\to\infty}|X_n-x_n|=0)=1 \tag{17}$$

对于 x_n 的期望计算如下:

$$E(x_n)=E\left(\beta\sum_{i=0}^{\infty}\alpha^i S_{n-i-1}\right)=\beta\sum_{i=0}^{\infty}\alpha^i E(S_{n-i-1})=\frac{\beta}{\lambda}\sum_{i=0}^{\infty}\alpha^i=\frac{\beta}{\lambda(1-\alpha)}=E \tag{18}$$

由于当 n 足够大时,X_n 会收敛到 x_n,因此当 n 足够大时,可以用 x_n 的期望去估计 X_n 的期望.如图 4 所示,使用图中的虚线来估计图中的实线表示的真实情况.

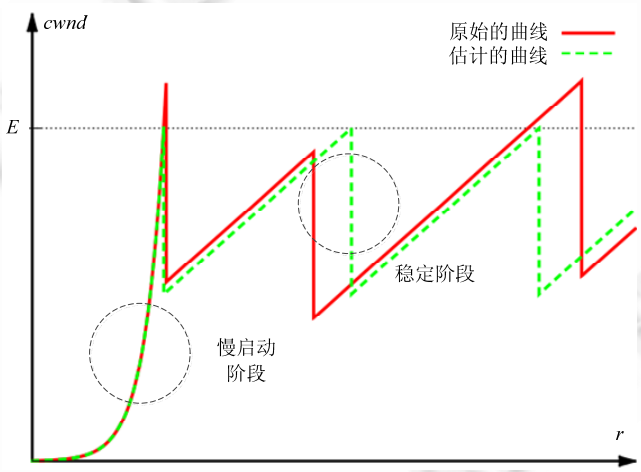


Fig.4 Sketching RTT estimation when W is larger compared with BDP

图 4 当 W 相对于 BDP 较大时 RTT 估计示意

虚线部分包含两个阶段:慢启动阶段和稳定阶段.与上一小节类似,慢启动经历的总轮次数为

$$R'_{SS}=\log_2 E \tag{19}$$

则慢启动阶段发送的报文数 N_{SS} 为

$$N_{SS}=\sum_{r=0}^{R_{SS}}2^r=2E-1 \tag{20}$$

稳定阶段时,可以计算出平均传输速率为

$$v=\bar{X}=\frac{E+\alpha E}{2}=\frac{\beta(1+\alpha)}{2\lambda(1-\alpha)} \tag{21}$$

其中,$\beta=k/\tau$.

稳定阶段发送的报文数 N_S 为

$$N_S = v\frac{T-R'_{SS}\tau}{\tau} = \frac{\beta(1+\alpha)}{\lambda(1-\alpha)}\left(\frac{T}{\tau}-\log_2 E\right) \tag{22}$$

其中,T 为传输总时间.

对于总报文数为 N 的 TCP 传输,有:

$$N = N_{SS}+N_S = 2E-1+\frac{\beta(1+\alpha)}{2\lambda(1-\alpha)}\left(\frac{T}{\tau}-\log_2 E\right) = \frac{\beta}{\lambda(1-\alpha)}+\frac{\beta(1+\alpha)}{2\lambda(1-\alpha)}\left(\frac{T}{\tau}-\log_2\frac{\beta}{\lambda(1-\alpha)}\right)-1 \tag{23}$$

将$\lambda=Np/T$,$\beta=k/\tau$带入公式(23),则可转化为关于 RTT 的估计值τ的方程,即可求得对 RTT 的估计.

由于上式较为复杂,在实际测量中,当用于计算的流记录的总报文数足够多时,可以忽略图 4 所示的慢启动阶段以达到简化的目的.此时,公式(23)可转化为

$$N = v\frac{T}{\tau} = \frac{\beta(1+\alpha)}{2\lambda(1-\alpha)}\frac{T}{\tau} = \frac{k}{2\tau}\frac{T}{Np}\frac{1+\alpha}{1-\alpha}\frac{T}{\tau} \tag{24}$$

则 RTT 的估计值τ为

$$\tau = \frac{T}{N}\sqrt{\frac{k(1+\alpha)}{2p(1-\alpha)}} \tag{25}$$

其中,T 为流持续时间,N 为总报文数,p 为丢包率,α,k 为拥塞控制策略中的相关常量.

3.3 当W与BDP相近时

当 W 与 BDP 相近时,*cwnd* 的增长将有可能达到 W 从而停止增长,因此有:

$$X_{n+1} = \min(\alpha X_n+\beta S_n, W) = \alpha X_n+\beta\min\left(S_n,\frac{W-\alpha X_n}{\beta}\right) \tag{26}$$

其中,运算 min(x,y)表示取 x,y 中较小的一个值,其他变量的定义如上一小节.又因为 X_n 的取值范围为[0,W],则有:

$$\alpha X_n+\beta\min\left(S_n,\frac{(1-\alpha)W}{\beta}\right) \leqslant X_{n+1} \leqslant \alpha X_n+\beta\min\left(S_n,\frac{W}{\beta}\right) \tag{27}$$

考虑随机过程$\{Y_n\}$定义如下:

$$Y_{n+1} = \alpha Y_n+\beta\min\left(S_n,\frac{(1-\alpha)W}{\beta}\right) \tag{28}$$

易知,随机等式(28)右边的后半部分同样是 i.i.d.的,并且期望的对数小于正无穷,因此同样满足文献[16]中的定理 1 的条件,从而有唯一的稳定解 y_n:

$$y_n = \beta\sum_{i=0}^{\infty}\alpha^i\min\left(S_{n-i-1},\frac{(1-\alpha)W}{\beta}\right) \tag{29}$$

满足:当 n 趋向于正无穷时,y_n 等于 Y_n 的概率为 1.

显然,若 $X_0=Y_0$,对于任意的自然数 n,有 $Y_n \leqslant X_n$,则:

$$E(x_n) \geqslant E(y_n) = \frac{\beta}{1-\alpha}E\left(\min\left(S_{n-i-1},\frac{(1-\alpha)W}{\beta}\right)\right) \tag{30}$$

又,对于任意的 n,S_n 为参数为λ的指数分布,因此,

$$E\left(\min\left(S_{n-i-1},\frac{(1-\alpha)W}{\beta}\right)\right) = \int_0^{\frac{(1-\alpha)W}{\beta}}x\lambda e^{-\lambda x}dx+\int_{\frac{(1-\alpha)W}{\beta}}^{\infty}\frac{(1-\alpha)W}{\beta}\lambda e^{-\lambda x}dx = \frac{1-e^{-\frac{(1-\alpha)W}{\beta}\lambda}}{\lambda} \tag{31}$$

因此有:

$$E(x_n) \geqslant E(y_n) = \frac{\beta}{\lambda(1-\alpha)}\left(1-e^{-\frac{(1-\alpha)W}{\beta}\lambda}\right) \tag{32}$$

同理,对于$\{Z_n\}$:

$$Z_{n+1}=\alpha Z_n+\beta\min\left(S_n,\frac{W}{\beta}\right) \tag{33}$$

进行同样的分析,可得:

$$E(x_n)\leqslant\frac{\beta}{\lambda(1-\alpha)}\left(1-\mathrm{e}^{-\frac{W}{\beta}\lambda}\right) \tag{34}$$

又,当忽略慢启动阶段时,有:

$$N=v\frac{T}{\tau}=\frac{(1+\alpha)E(x_n)}{2}\frac{T}{\tau},$$

则可得到不等式:

$$\frac{(1+\alpha)kT^2}{2Np(1-\alpha)\tau^2}\left(1-\mathrm{e}^{-\frac{(1-\alpha)WNp\tau}{kT}}\right)\leqslant N\leqslant\frac{(1+\alpha)kT^2}{2Np(1-\alpha)\tau^2}\left(1-\mathrm{e}^{-\frac{WNp\tau}{kT}}\right) \tag{35}$$

考察函数$f(x)$:

$$f(x)=b\frac{1-\mathrm{e}^{-ax}}{x^2} \tag{36}$$

其中,$a>0,b>0,x>0$,$f(x)$的导数为

$$f'(x)=b\frac{\mathrm{e}^{-ax}(ax+2-2\mathrm{e}^{ax})}{x^3} \tag{37}$$

考虑函数$g(x)=2\mathrm{e}^{ax}-ax-2$,$g(x)$的导数为$g'(x)=2a\mathrm{e}^{ax}-a$,当$x\geqslant 0$且$a>0$时,则有$g'(x)>0$.因此,当$x>0$且$a>0$时,有:$g(x)>g(0)=0$,即$ax+2-2\mathrm{e}^{ax}<0$,即$f'(x)<0$.因此,$f(x)$在$x>0$上是单调递减函数.

由以上分析可知:对于关于τ的不等式方程(35),有且只有一个解:

$$\tau_1\leqslant\tau\leqslant\tau_2 \tag{38}$$

其中,τ_1,τ_2可以通过如下两个方程求得:

$$\begin{cases}\dfrac{(1+\alpha)kT^2}{2Np(1-\alpha)\tau_1^2}\left(1-\mathrm{e}^{-\frac{(1-\alpha)WNp}{kT}\tau_1}\right)=N\\ \dfrac{(1+\alpha)kT^2}{2Np(1-\alpha)\tau_2^2}\left(1-\mathrm{e}^{-\frac{WNp}{kT}\tau_2}\right)=N\end{cases} \tag{39}$$

其中,W为发送端 Socket 缓冲区大小,T为流持续时间,N为总报文数,p为丢包率,α,k为拥塞控制策略中的相关常量.

3.4 抽样下的讨论

上述 3 种情况下的 RTT 估计只是涉及到一个 TCP 流的持续时间 T 与总报文数 N 两个变量,在抽样环境中,关于已知抽样后的报文数 N_S,获得实际传输中的报文数 N 可以通过如下方法[17,18]来估计:

对于一个流记录,令 X 表示原始报文数,Y 表示抽样后的报文数,假设对于每个报文,被抽到的事件相互独立,且概率为一个定值(抽样比 s),则 $P_x=P(X=x|Y=N_S)$的计算如下所示:

$$P_x=P(X=x\mid Y=N_S)=\frac{P(Y=N_S\mid X=x)P(X=x)}{P(Y=N_S)}=\frac{P(Y=N_S\mid X=x)P(X=x)}{\sum_{x'=N_S}^{\infty}P(Y=N_S\mid X=x')P(X=x')} \tag{40}$$

研究表明[17],网络中的流服从重尾分布,故使用 Pareto 分布来估计原始流长的分布,并且使用参数为 1 的 Pareto 分布来估计未知原始流分布中的流长分布[18],即$\beta=1$,且 $X_{\min}$ 为 N_S,则有:

$$P(X=x)=\begin{cases}0, & x<N_S\\ \dfrac{\beta N_S^{\beta}}{x^{\beta+1}}=\dfrac{N_S}{x^2}, & x\geqslant N_S\end{cases} \tag{41}$$

则原始的报文数 N 的估计为

$$N=\sum_{x=N_S}^{\infty}xP_x=\sum_{x=N_S}^{\infty}x\frac{P(Y=N_S\mid X=x)P(X=x)}{\sum_{x'=N_S}^{\infty}P(Y=N_S\mid X=x')P(X=x')}$$
$$=\sum_{x=N_S}^{\infty}x\frac{\binom{x}{N_S}s^{N_S}(1-s)^{x-N_S}\frac{N_S}{x^2}}{\sum_{x'=N_S}^{\infty}\binom{x'}{N_S}s^{N_S}(1-s)^{x'-N_S}\frac{N_S}{x'^2}}=\frac{\sum_{x=N_S}^{\infty}\frac{1}{x}\binom{x}{N_S}(1-s)^x}{\sum_{x'=N_S}^{\infty}\frac{1}{x'^2}\binom{x'}{N_S}(1-s)^{x'}} \tag{42}$$

其中,s 为抽样比,N_S 为抽样后的报文数.虽然上式中分子和分母的展开式有无穷多项,但是由于越往后的项值越小,因此可以根据精度要求计算前几项.而若知道抽样后的持续时间 T_S,则可用 T_S 粗略地估计原始持续时间 T.

通过以上讨论可知,本节讨论的各种方法在抽样环境中仍然可以使用.

4 模型评估

4.1 仿真评估

本文使用 ns-2[19]进行仿真实验,采用的 TCP 版本是 TCP NewReno[20],此时,α为 0.5,k 为 1.分别对第 3 节中的 3 种情况进行仿真,每种参数下采用多次(50 次)实验取平均的方法,以提高实验结果的可信度.

4.1.1 当 W 相对于 BDP 较小时的仿真

图 5 是 ns-2 的仿真结果,仿真的基本参数:MSS 为 1 024,带宽为 1Gb/s,传输持续时间为 200s,RTT 为 200ms,TCP 窗口(模拟套接字缓冲区大小)为 100 个报文,丢包率为 0.01%.

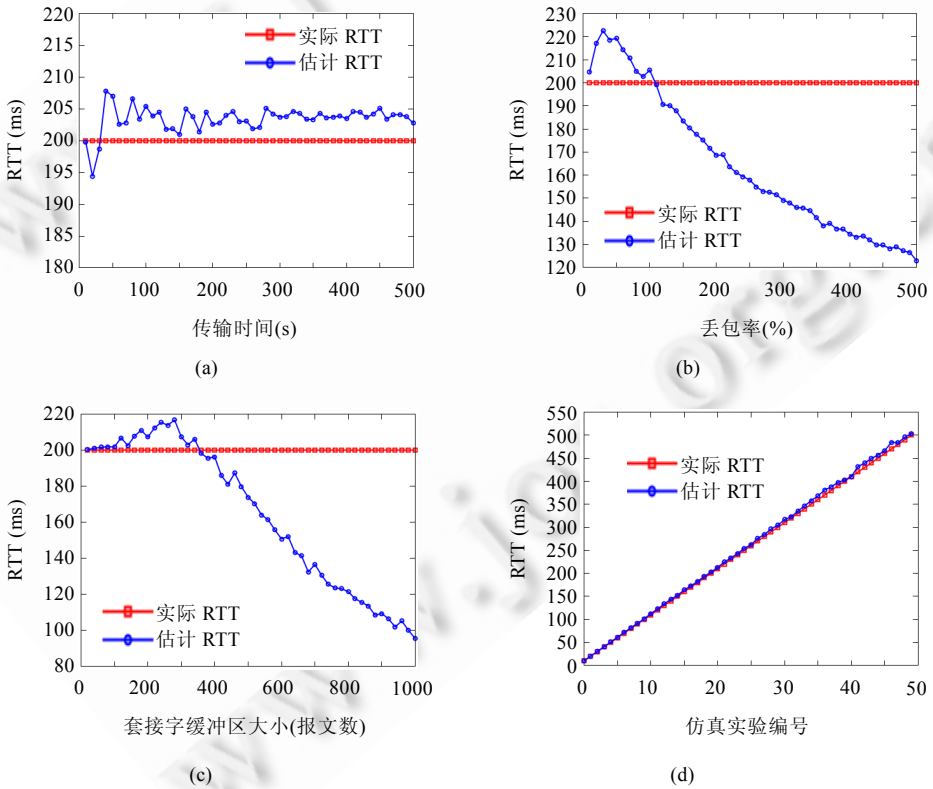


Fig.5 Simulation result when W is small compared with BDP

图 5 当 W 相对于 BDP 较小时的仿真结果

由于 ns-2 中的带宽仅在计算延迟时才使用,并不影响 TCP 传输,因此在仿真中,使用丢包率来模拟真实情况下的 BDP.这是因为,在真实传输中,当 BDP 较小时,必然丢包率较大;反之也成立.实验中,使用公式(11)对 RTT 进行估计.

图 5(a)为改变传输时间进行仿真时,传输时间与估计结果的关系.可以发现,传输时间对估计结果影响不大.图 5(b)为改变丢包率进行仿真时,丢包率与估计结果的关系.可以发现:随着丢包率的增大,估计的结果越来越差.这是因为当丢包率过大时,假设"每次丢包的恢复阶段直至 *cwnd* 增长为 W 之前不会发生新的丢包"在大部分情况下不再成立,此时,真实的 S_{loss} 比公式(4)中估计的 $N_{loss}\times S_{CA}$ 要小很多,故导致 RTT 估计值偏小.图 5(c)为改变 TCP 窗口进行仿真时,TCP 窗口与估计结果的关系.可以发现:随着 TCP 窗口的增大,估计的结果越来越差.与图 5(b)的情况类似,当 TCP 窗口,即套接字缓冲区大小增大时,假设在大部分情况下不再成立,此时,真实的 S_{loss} 比公式(4)中估计的 $N_{loss}\times S_{CA}$ 要小很多,故导致 RTT 估计值偏小.图 5(d)为改变真实的 RTT 进行仿真时,真实的 RTT 与估计结果的关系.可以发现:真实的 RTT 对估计结果影响不大.

从以上结果可知:当丢包率较小,同时,套接字缓冲区大小较小时,估计结果较为准确,这正符合 W 相对于 BDP 较小时的情形.

4.1.2 当 W 相对于 BDP 较大时的仿真

图 6 是 ns-2 的仿真结果,仿真的基本参数:MSS 为 1 024,带宽为 1Gb/s,传输持续时间为 200s,RTT 为 200ms,TCP 窗口(模拟套接字缓冲区大小)为 1 000 个报文,丢包率为 1%.实验中,使用公式(25)对 RTT 进行估计.

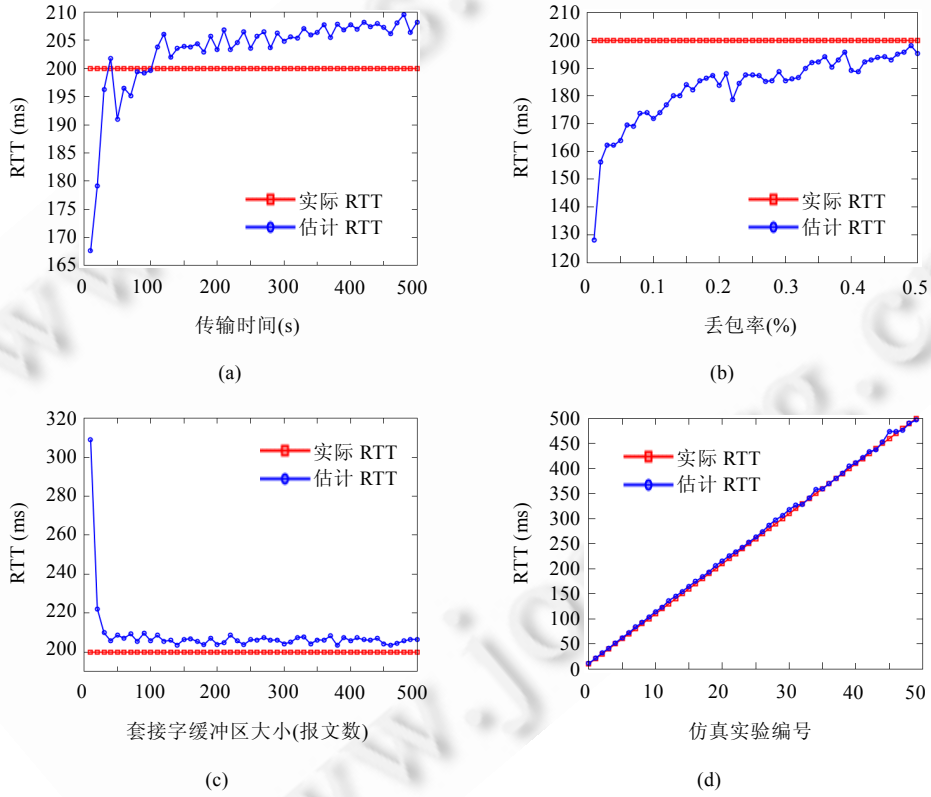


Fig.6 Simulation result when W is large compared with BDP

图 6 当 W 相对于 BDP 较大时的仿真结果

图 6(a)为改变传输时间进行仿真时,传输时间与估计结果的关系.可以发现:传输时间较小时,估计结果不理想.这是因为估计时忽略了慢启动过程,当传输时间较短时,将会带来较大的误差.同时,根据第 3.2 节中的讨论,

必须在 n 较大,即传输时间较长的情况下,才能用 $E(x_n)$去估计 $E(X_n)$,因此导致结果误差较大;而当传输时间较大时(>20s),传输时间对估计结果影响不大.图 6(b)为改变丢包率进行仿真时,丢包率与估计结果的关系.可以发现:随着丢包率的增大,估计的结果越来越好.这是因为,当丢包率较小时,发生丢包的次数较少,而使用公式(25)对 RTT 进行估计,是建立在随机过程的模型上的,因此估计结果变化剧烈.图 6(c)为改变 TCP 窗口进行仿真时,TCP 窗口与估计结果的关系.可以看到:当 TCP 窗口较小时,估计的结果较差.这是因为,当 TCP 窗口较小时,*cwnd* 的增长迅速地达到了 TCP 窗口,此时,*cwnd* 不再增长;而在估计中假设不会出现这种情况,因此导致估计的平均传输速率 v 偏大.根据公式(24),导致 RTT 的估计值偏大.图 6(d)为改变真实的 RTT 进行仿真时,真实的 RTT 与估计结果的关系.可以发现,真实的 RTT 对估计结果影响不大.

从以上结果可知:当丢包率较大,并且套接字缓冲区大小较大时,估计结果较为准确.这正符合 W 相对于 BDP 较大时的情形.同时,如果传输时间较短,不可忽略慢启动的影响,并且在丢包率特别低的情况下估计效果不好.但也应注意到,实际传输中几乎不会存在 BDP 较小、同时丢包率也较小的情况.因此,这种方法在实际情况下能得到较好的结果.

图 7 为一个典型情况的示意,可以发现:虽然在单一时间点上误差较大,但是总体来看,误差很小.

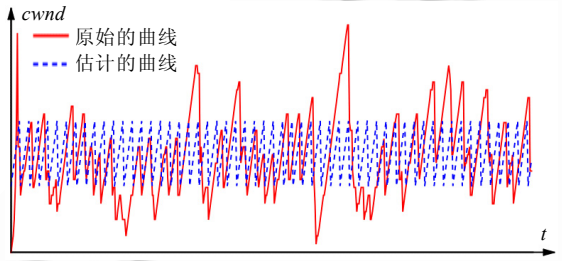


Fig.7 Sketching simulation when W is large compared with BDP

图 7 W 相对于 BDP 较大时的仿真示意

4.1.3 当 W 与 BDP 相近时的仿真

取仿真的基本参数:MSS 为 1 024,带宽为 1Gb/s,传输持续时间为 200s,RTT 为 200ms,TCP 窗口(模拟套接字缓冲区大小)为 50 个报文进行仿真,变化丢包率从 0.02%~1%,共 50 组.其中,不等式方程(38)是通过牛顿迭代法(Newton's method)[21]来计算的,结果精确到 1ms.最终结果如图 8 所示.

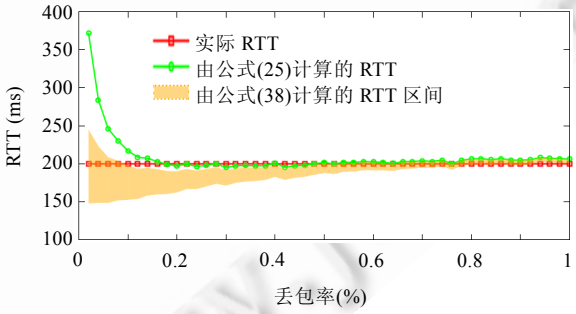


Fig.8 Simulation result when W is comparable to BDP

图 8 当 W 与 BDP 相近时的仿真结果

由图 8 可知:当丢包率较小时,通过公式(25)计算的 RTT 明显偏大,但是通过公式(38)计算的 RTT 区间仍能有效地估计 RTT;而当丢包率增大后,公式(38)计算的 RTT 区间与公式(25)计算的 RTT 趋于一致.

4.1.4 对于抽样的仿真

取仿真的基本参数:MSS 为 1 024,带宽为 1Gb/s,丢包率为 0.5%,RTT 为 200ms,TCP 窗口(模拟套接字缓冲区大小)为 1 000 个报文进行仿真,变化传输持续时间从 20s~1000s,共 50 组.分别采用抽样比 1,1/4,1/16,1/64,1/256 进行结果对比,如图 9 所示.

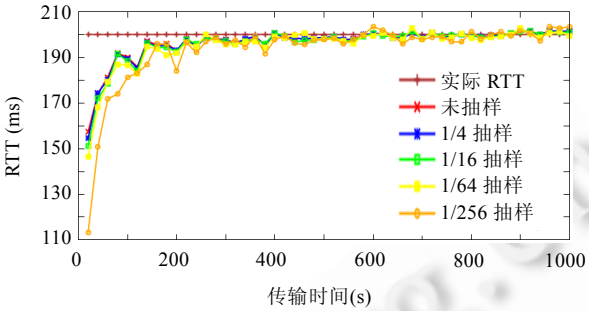


Fig.9 Simulation result for sampled flow data

图 9 对抽样流的仿真结果

从图中可以看出:随着传输时间的增长,抽样后的结果与未抽样的结果越来越接近.因此,对于传输时间较长的抽样流记录,本文方法的精度同样较高.

4.2 实测数据评估

由第 4.1 节可以看出:在仿真实验中,第 3 节提出的模型能够很好地估计 RTT.为了进一步测试估计效果,本文使用实际数据估计,并使用 SA 算法[2]和文献[6]提出的 PRE 算法作为对比数据.

实验用来自于 IPTAS 系统[22]采集的两个 CERNET 节点间 3 个小时的实际 trace 组成的流,并且筛选出源主机运行 Window 系统(TCP 版本为 TCP NewReno)的、源端口为 80(HTTP)和 20(FTP 数据)的 TCP 流.同时,使用一个流中的报文的 TCP 序列号的错序次数来估计这个流传输过程中的丢包数,进而计算丢包率.为了简化实验,当丢包率较小时,采用第 3.1 节中的方法估计 RTT;而当丢包率较大时,采用第 3.2 节中的方法估计 RTT.

为方便统计与结果对比,本文将流长(流内报文数)以 100 为间隔分组,第 i 个区间的流长范围为[100×(i–1), 100×i].注意到:对于 PRE 算法,在流长较小的情况下不能计算;同时,在采集到的数据中,当流长大于 2 500 时,每组的流数较小(小于 100),结果意义不大.因此,只筛选出流长大于 10 且小于 2 500 的流,分为 25 个流长区间(第 1 个区间的范围为[10,100]).在这个 trace 中,符合条件的有 96 190 个流.具体分布如图 10 所示.

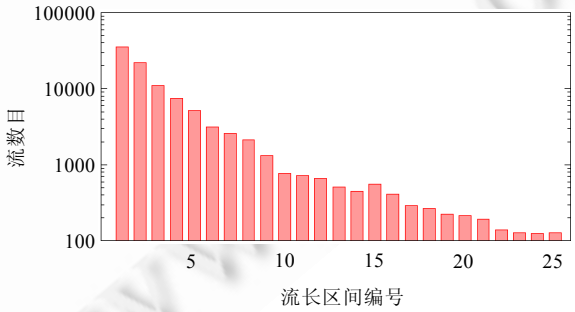


Fig.10 Flow length distribution

图 10 流长分布

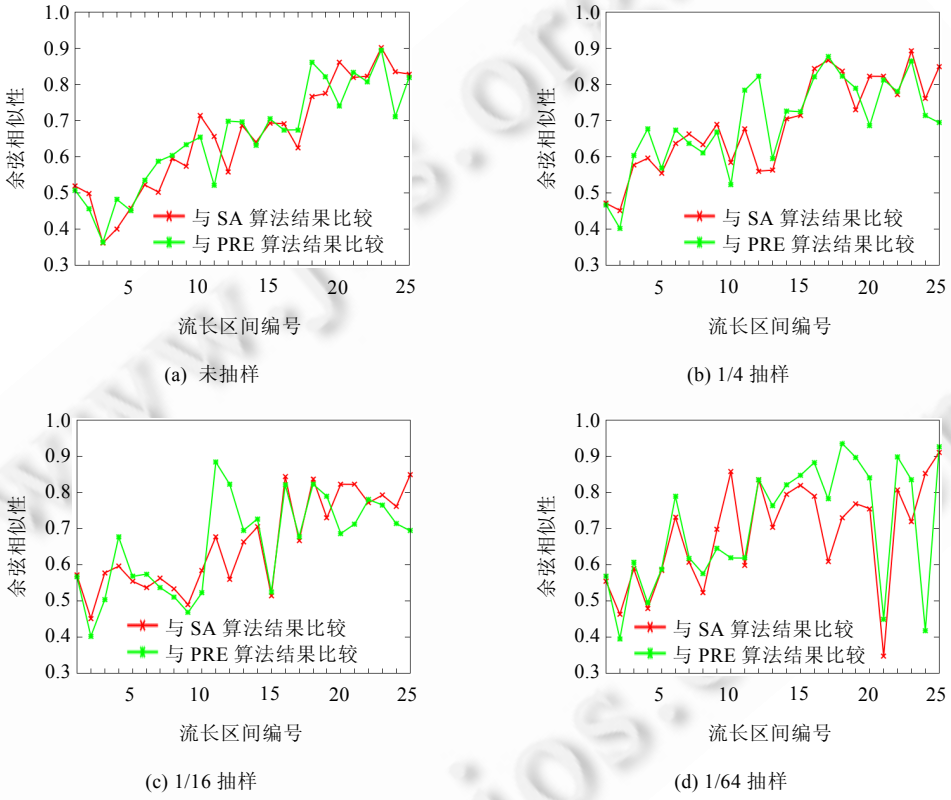
为了测试抽样对本文提出算法的影响,我们使用抽样比分别为 1(未抽样),1/4,1/16,1/64 的流记录来计算

RTT 估计值.由于 SA 算法和 PRE 算法均不支持抽样,所以采用这些抽样的流记录对应的未抽样的流记录中的报文序列作为这两种算法的输入,计算对比数据.注意:在抽样情况下,数据分组的标准仍然是原始的流长而不是抽样后的流长.

由于每一组中的计算结果是对不同主机间的 RTT 的估计结果,而基于统计的误差分析是对同一实验的多次实验结果进行分析,因此在这里并不适用.为了衡量本文提出的算法与 SA 算法和 PRE 算法的结果的差异性,使用余弦相似性(cosine similarity)[23]进行计算,取值从–1 到 1,越接近 1,代表两组数据越相似.对于两组数据 $A=\{a_i\}_n$,$B=\{b_i\}_n$.它们的余弦相似性 $S_{\cos}$ 的定义为

$$S_{\cos}=\frac{\sum_n a_i b_i}{\sqrt{\sum_n a_i^2}\sqrt{\sum_n b_i^2}} \tag{43}$$

图 11 所示的是在 4 种抽样比下,将所有流按上述的流长区间分组,并分别计算各分组内的估计结果与 SA 算法和 PRE 算法的余弦相似性的结果.



(a) 未抽样 (b) 1/4 抽样

(c) 1/16 抽样 (d) 1/64 抽样

Fig.11 Experiment results of real flow data

图 11 实际数据的实验结果

可以发现:随着流长越来越长,估计的效果也越来越好.但是随着抽样比的增大,结果的波动性也在增大.这主要是由于在抽样比较大的情况下,对于流长估计的偏差比较大引起的.但是通过第 3.4 节的讨论可知,估计结果的期望并不会受到抽样的影响.因此,如果能在一定时间内对特定主机间的足够多的抽样流记录进行估计,并对结果取平均值,可以有效地消除高抽样比带来的结果的波动性.为了更进一步地证明这个结论,本文在上述条件相同的情况下,分别对 3 小时、6 小时和 9 小时的 trace 进行 1/64 抽样的实验,结果只采用 PRE 算法进行对比,结果如图 12 所示.可以发现:当数据增加后,确实能够有效地降低估计结果的波动性.

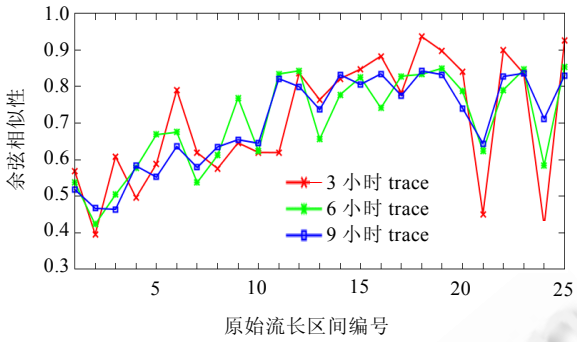


Fig.12 The influence of the amount of data on estimation result

图 12 数据的数量对估计结果的影响

5 总 结

往返时延是刻画网络性能的一项重要指标.一个 RTT 的值是对特定两个 IP 地址之间的传输性能的重要衡量标准之一.传统的 RTT 测量方法不是基于全报文的就是基于未抽样的流记录,随着网络的规模越来越大,对于大规模主干网环境下,即使采集未抽样的流记录也变得越来越困难.

为了适应大规模主干网环境下的 RTT 测量,本文提出一种基于抽样流记录的 RTT 测量方法,通过对 AIMD 型的 TCP 块状流的传输特性进行分析,分别建立了以下 3 种情况下的 RTT 估计模型:(1) 当套接字缓冲区大小相对于 BDP 较小时;(2) 当套接字缓冲区大小相对于 BDP 较大时;(3) 当套接字缓冲区大小与 BDP 相近时.

实验结果表明:在 3 种不同的情况下,各自的模型都能较好地完成 RTT 估计.在带宽较大、丢包率较低的情况下,使用第 3.1 节中的方法进行计算 RTT 更精确;而在带宽较小、丢包率较高的情况下,使用第 3.2 节中的方法计算 RTT 更精确;当链路情况介于两者之间时,可以使用第 3.3 节介绍的方法来估计出 RTT 区间.由于在估计当中只使用了流持续时间和总报文两个变量,因此,本方法同样适用于以抽样流记录为输入的环境.

在实际网络测量中,TCP 块状流的发送端通常是服务器(Web 服务器或 FTP 服务器),所以套接字缓冲区大小几乎不变;同时,主机间的网络状态也相对固定,因此一旦确定了测量目的,通过一次性收集相关的参数,即可判断出适用每个端系统对的计算模型,很少情况下会由于套接字缓冲区大小与 BDP 的对比发生改变而需要对模型进行更换.另一方面,虽然由于只有块状流记录能作为测量的源数据可能会导致测量需求中的某些主机不能被测量到,但是分别与要测量的两个主机在相似的网络拓扑结构(如分别在相同子网中)的两个主机间的测量所得数据,能在一定程度上代表未被测量到的主机的网络性能.本文所设计的 RTT 测量方法适用于在主干网中对网际的时延进行持续地观测,以满足网络运行管理系统日常网络服务质量监测的需要.它可以替代传统的使用大规模主动测量平台的方法,避免周期性主动测量的开销.在进行性能诊断时,本文方法与传统的主动性能测量方法构成互补,网络管理人员可以在这种宏观测量的基础上进行有针对性的主动测量来获得进一步的诊断信息.

通过分析仿真数据和实际数据进行实验可以发现:无论输入数据是未抽样的流记录还是抽样的流记录,估计的准确性均随着流长的增长而有所提高.但是,随着抽样比的增大,结果的波动性也增大.抽样并不会影响结果的期望,故当有足够多的数据时,就可以在高抽样比的情况下估计 RTT.因此,在有大量数据的大规模骨干网络情况下,本文提出的方法在抽样的环境中可以具有很好的实用性.

References:

[1] Stevens WR. TCP/IP Illustrated, Vol.1: The Protocol. Reading: Addison Wesley, 1993.

[2] Jiang H, Dovrolis C. Passive estimation of TCP round-trip Times. In: Proc. of the SIGCOMM 2002. 2002. 75−88. [doi: 10.1145/571697.571725]

[3] Veal B, Li K, Lowenthal D. New methods for passive estimation of TCP round-trip times. In: Proc. of the PAM 2005. 2005. 121−134. [doi: 10.1007/978-3-540-31966-5_10]

[4] Zhang Y, Breslau L, Paxson V, Shenker S. On the characteristics and origins of Internet flow rates. In: Proc. of the SIGCOMM 2002. 2002. 309−322. [doi: 10.1145/633025.633055]
[5] Jaiswal S, Iannaccone G, Diot C, Kurose J, Towsley D. Inferring TCP connection characteristics through passive measurements. In: Proc. of the INFOCOM 2004. 2004. 1582−1592. [doi: 10.1109/INFCOM.2004.1354571]
[6] Zhang YB, Lei ZM. A passive RTT estimate algorithm for TCP. Journal of Beijing University of Posts and Telecommunications, 2004,27(5):85−89 (in Chinese with English abstract). [doi: 10.3969/j.issn.1007-5321.2004.05.017]
[7] Claise B. Cisco systems NetFlow services export version 9. RFC 3954, 2004.
[8] sFlow. http://www.sflow.org/index.php
[9] Claise B. Specification of the IP flow information export (IPFIX) protocol for the exchange of IP traffic flow information. RFC 5101, 2008.
[10] Strohmeier F, Dorfinger P, Trammell B. Network performance evaluation based on flow data. In: Proc. of the IWCMC 2011. 2011. 1585−1589. [doi: 10.1109/IWCMC.2011.5982608]
[11] Altman E, Avrachenkov K, Barakat C. A stochastic model of TCP/IP with stationary random losses. In: Proc. of the SIGCOMM 2000. 2000. 231−242. [doi: 10.1145/347059.347549]
[12] Tomita N, Valaee S. Data uploading time estimation for CUBIC TCP in long distance networks. Computer Networks, 2012,56(11): 2677−2689. [doi: 10.1016/j.comnet.2012.04.010]
[13] Bao W, Wong VWS, Leung VCM. A model for steady state throughput of TCP CUBIC. In: Proc. of the Global Telecommuni-cations Conf. (GLOBECOM 2010). 2010. 1−6. [doi: 10.1109/GLOCOM.2010.5684172]
[14] Allman M, Paxson V, Blanton E. TCP congestion control. RFC 5681, 2009. http://tools.ietf.org/html/rfc5681
[15] Karlin S, Taylor HM. A First Course in Stochastic Processed. 2nd ed., Singapore: Elsevier Pte Ltd., 2007.
[16] Brandt A. The stochastic equation $Y_{n+1}=A_nY_n+B_n$ with stationary coefficients. In: Advances in Applied Probability. 1986. 211−220. http://www.jstor.org/discover/10.2307/1427243
[17] Duffield N, Lund C, Thorup M. Estimating flow distributions from sampled flow statistics. In: Proc. of the SIGCOMM 2003. 2003. 325−336. [doi: 10.1145/863955.863992]
[18] Zhang XY, Gong J, Wu H. A method of estimating average round-trip latency based on specific flow records in NetFlow. Computer Applications and Software, 2010,27(5):64−67 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-386X.2010.05.020]
[19] McCanne S. The network simulator—ns-2. 1997. http://www.isi.edu/nsnam/ns/
[20] Floyd S, Henderson T, Gurtov A. The NewReno modification to TCP's fast recovery algorithm. RFC 3782, 2004. http://tools.ietf.org/html/rfc3782
[21] Wikipedia. Newton's method. 2013. http://en.wikipedia.org/wiki/Newton%27s_method
[22] Jiangsu Key Laboratory of Computer Networking Technology. IP trace distribution system (IPTAS). 2013. http://iptas.edu.cn/src/system.php
[23] Wikipedia. Cosine similarity. 2013. http://en.wikipedia.org/wiki/Cosine_similarity

附中文参考文献:

[6] 张轶博,雷振明.一种被动式 RTT 测量算法.北京邮电大学学报,2004,27(5):85−89. [doi: 10.3969/j.issn.1007-5321.2004.05.017]
[18] 张晓宇,龚俭,吴桦.一种基于 NetFlow 特定流记录的平均往返时延估计方法.计算机应用与软件,2010,27(5):64−67. [doi: 10.3969/j.issn.1000-386X.2010.05.020]

苏琪(1989－),男,江西上饶人,博士生,主要研究领域为网络管理,网络测量.
E-mail: qsu@njnet.edu.cn

苏艳珺(1988－),女,硕士生,主要研究领域为网络管理,网络测量.
E-mail: wangzhejunzi@163.com

龚俭(1957－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络管理,网络安全.
E-mail: jgong@njnet.edu.cn