[1]TCP 流报文往返比研究

陈兴其，丁伟，吴桦，龚俭
(东南大学计算机科学与工程学院；江苏省计算机网络技术重点实验室 南京 210096)

摘要：测度是网络测量的基础。本文提出了 TCP 流报文往返比这个测度。通过对 CERNET 主干网采集的 trace 进行组流分析，从 TCP 双向流个数随报文往返比的分布情况得出：经过本文所分析主干网链路的 TCP 流报文往返比集中在 1.2。并从 TCP 协议角度分析了影响这个测度的决定因素——TCP 协议的 Delayed ACK 应答机制。
关键字：报文往返比，测度，组流，网络测量

Study on TCP Flow Bidirectional Packet Ratio

CHEN Xing-qi,DING Wei,WU Hua,GONG Jian
（College of Computer Science & Engineering, Southeast University; Key Laboratory of Computer Network Technology in Jiangsu, Nanjing 210096, China）

Abstract: Metric is the basis of network measurement. A new metric was proposed in this paper as BPR(Bidirectional Packet Ratio) of TCP flow. After analysis of aggregate flow with the trace from CERNET backbone and together with the distribution of TCP flows and BPR, we came to this conclusion that TCP flow BPR concentrates at 1.2. Finally, from the view of TCP protocol, it's the "Delayed ACK" mechanism that affects this metric.
Keywords: BPR, metric, aggregate flow, network measurement

中图分类号：TP393.0　　**文献标识码：A**

1 引言

对互联网特性以及用户行为的研究主要是通过网络测量和网络流量分析来进行的。网络测量的结果由测度体现出来，它对于网络测量具有很重要的意义。测度在 RFC2330 中被定义为“在可操作的互联网中，有一些衡量网络性能以及可靠性的参数。这些参数值是我们所希望得到的，当这些参数被精确描述后，我们将其定义为测度”[2]。IETF 规定测度必须是具体的并且定义明确；用于测量测度的算法要具有可重复性；测度必须有用。

本文提出了 TCP 报文往返比这个测度并研究了它在网络测量中的影响。通过对 CERNET 主干江苏省网边界采集到的 trace 进行组流实验分析得出结论：经过此采集点的 TCP 流报文往返比集中在 1.2。由于报文往返比受 TCP 协议内部应答机制的影响。所以本文将从研究 TCP 协议的应答机制入手来寻找影响报文往返的因素及规律。

2 TCP 应答机制研究

在简单的 TCP 交互过程中，接收方在收到一个报文后会向数据发送方应答一个 ACK 报文。如果在大量数据传输过程情况下，接收到一个数据报文立即应答一个 ACK，那么在网

[1]本文受国家科技支撑计划(No. 2008BAH37B04)和国家 973 计划 (No. 2009CB320505)资助。
作者简介：陈兴其（1986-），男，硕士研究生；丁伟（1962-），女，教授。
E-Mail: xqchen@njnet.edu.cn

络中将会出现大量的 ACK 报文。这极大的影响了互联网上路由器及主机效率。所以当前主流操作系统在 TCP 协议实现中引入了累计确认机制，即下面将要介绍的 Delayed ACK 机制。

接收 TCP 流数据的主机在平均收到一个报文后，可以通过发送少于一个的 ACK 报文来提高互联网及主机的效率[1]。现有的操作系统在 TCP 报文响应实现中，普遍使用 RFC1122 中提出的 Delayed ACK 机制，如 Windows 及 Linux。RFC1122 指出：“一个 TCP 实现应该包括累计确认机制，但是 ACK 报文不能无限期的被延迟发送，特别地，发送时延不能超过 0.5 秒并且在收到第二个‘全段’报文后应答一个 ACK 报文”[1]。 “全段”报文是指 IP 报文中数据长度为已方 MSS 的报文（在 TCP 链接建立初期双方相互通告自己的 MSS）。但是由于链路 MTU 可能小于 MSS，所以接收方接收到的最大报文数据长度可能小于已方 MSS。于是要接收 3 个及以上的报文才能达到两个“全段”报文的大小。这就是 RFC2525 里讨论的“Stretch ACK violation”问题。所以一般操作系统内核在 Delayed ACK 机制上是按照“收到第二个报文即回复一个 ACK”方式来实现的，比如 Windows。

在使用了 Delayed ACK 机制后，数据接收方在以下几种情况下发送 ACK 报文：

（1） 超时前收到第二个没有应答的报文；

（2） delayed ACK 超时（RFC1122 提出时延上限为 0.5 秒，绝大部分实现采用 0.2 秒）；

（3） 收到乱序报文；

（4） 发现丢包后再次收到报文。

3 测度定义和测量方法

在一次 TCP 连接中，设主机发送报文数 Ps，接收报文数 Pr。将报文往返比定义为数据发送方发送报文数与接收报文数之比，即 $\frac{Ps}{\mathrm{Pr}}$。定义主机平均报文长度为 $\frac{\sum_i PLi}{Ps}$，其中 PLi 为主机发送的第 i 个报文的长度，且认为平均报文长度值大者为数据发送方。

本实验采用被动测量方法，源数据为 CERNET 江苏省网到国家主干链路上 Watcher 系统采集的 trace。Watcher 为自行开发的运行在边界路由器上进行原始报文采集的系统。可以通过配置该系统来进行流量过滤。本实验所用到的数据是未经过滤的全报文采集到的 trace。

首先对采集到的 trace 按照五元组+16 秒超时[3]规则进行组流（单向流）。然后将两个方向（进出路由器）上的单向流按匹配规则合并成双向流记录（匹配规则为五元组+流起始时戳）。双向流记录是一次 TCP 连接的信息摘要，它包括了双向报文数、字节数以及五元组等信息。所以，可根据双向流记录直接计算其报文往返比。作为对比，实验采用了 Watcher 在 2005 年 11 月，2006 年 12 月以及 2008 年 12 月三个时间点采集的 30 分钟 trace。并分析了不同时段 trace 中 TCP 流个数随报文往返比的分布情况。更进一步地，我们对 TCP 流进行过滤，用同样的方法对 http 流（端口 80）以及 ftp 交互流（端口 21）进行报文往返比分析。

4 实验结果及分析

4.1 实验结果

实验结果用二维曲线表示：横坐标为报文往返比区间；纵坐标为落入相应报文往返比区间的 TCP 双向流数目。其中，横坐标每个刻度表示往返比区间长度 0.3，即刻度 1,2,3 分别表示报文往返比(0,0.3],(0.3,0.6],(0.6,0.9]。图 1,2,3 为不同时间的 trace 所对应双向流分布：

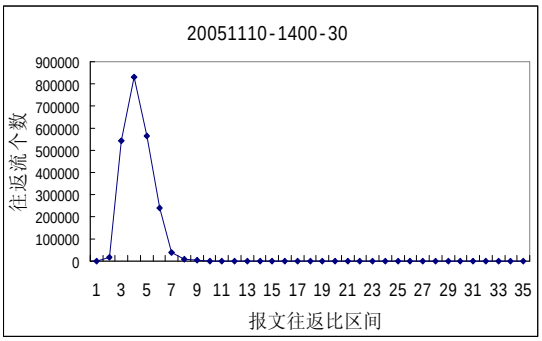

图 1 2005 年 11 月 10 日 TCP 流随报文往返比分布

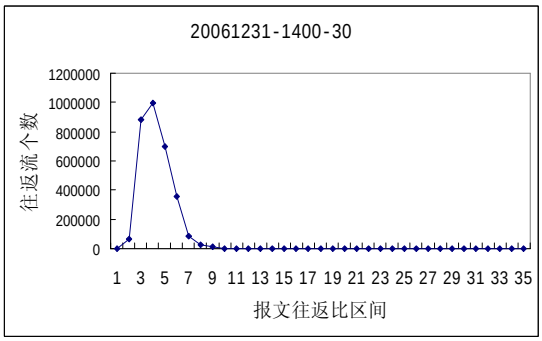

图 2 2006 年 12 月 31 日 TCP 流随报文往返比分布

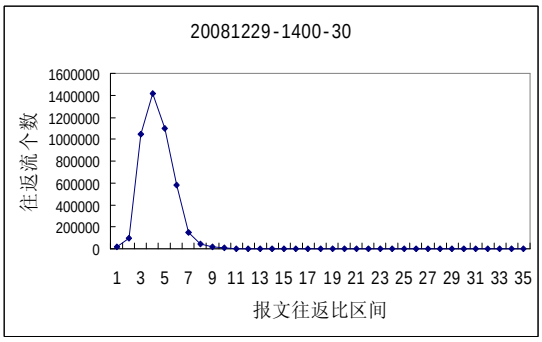

图 3 2008 年 12 月 29 日 TCP 流随报文往返比分布

图 4,5 为不同网络应用所对应双向流分布：

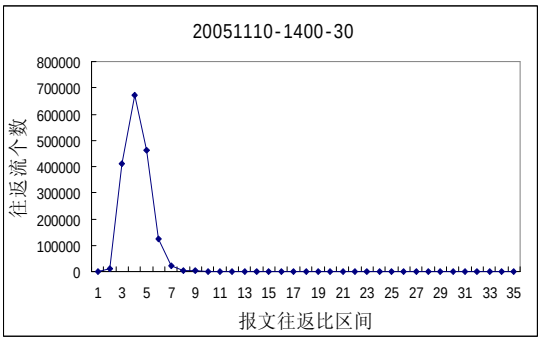

图 4 HTTP 流随报文往返比分布

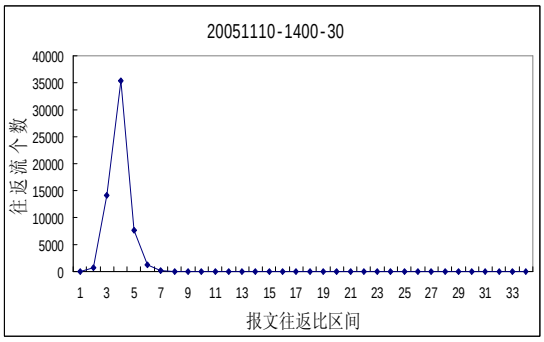

图 5 FTP 流随报文往返比分布

4.2 实验结果分析

从实验结果可以明显的看出曲线在刻度 4 处有一个极值点。刻度 4 表示报文往返比第四个区间。由于每个区间长度为 0.3，所对应的报文往返比为 4 * 0.3 = 1.2。

从 TCP 协议角度来看，在数据传输过程中，数据接收方在收到两个报文后向发送方应答一个 ACK 报文。如果网络状况良好（即不出现丢包情况以及网络延迟相对较小），大量的数据传输会使报文往返比值逼近 2。但是互联网是一个复杂的环境，一些情况的出现会使往返比值逼近 1。如：出现了 2 小节所述的第 2 种情况，累积超时结束将立即发送 ACK 报文。还有，在第 3 以及第 4 种情况以及在一些 TCP 实现中的快速模式（Linux 的 Quick ACK Mode）下，数据接收方每收到一个报文则立即应答 ACK 报文。综上所述，报文往返比值会在比值 1 和 2 之间，所以实验得出的结论——TCP 流报文往返比值主要集中在 1.2 是可接受的。而且，从以上实验结果可看出：不同时间段以及不同网络应用的流得出的结果结论是一致的。充分说明了 TCP 协议的 Delayed ACK 机制是影响报文往返比的决定因素。

5 结论

本文提出 TCP 流报文往返比这个测度，通过以上实验结果分析提出互联网中 TCP 报文往返比集中在 1.2 这个结论。并从 TCP 协议的 Delayed ACK 机制分析了这个结论的合理性。由于在固定点不同时间段采集到的 trace 以及在不同网络应用下这个结论有较好的稳定性。通过对这个值的观察，可以对网络的运行情况有初步的诊断。由于实验所进行的环境为 CERNET 华东（北）节点，结论的普遍性还需进一步验证。

参考文献：

[1] Braden.R, Editor, Requirements for Internet Hosts -- Communication Layers[P], RFC 1122, October 1989.

[2] V. Paxson, Framework for IP Performance Metrics[P], RFC 2330，May 1998

[3] 王远,丁伟,龚俭.TCP 数据流超时研究[J].厦门大学学报(自然科学版), 2007, Vol.42: 192-195.