第 26 卷第 7 期 电 子 与 信 息 学 报 Vol.26No.7
2004 年 7 月 Journal of Electronics & Information Technology July 2004

一种新的基于事件触发的网络流量采样测量方法及其应用研究 [1]

杨新宇 魏恒义 程竹林

(西安交通大学计算机科学与技术系 西安 710049)

摘 要: 该文主要对采样测量方法进行了研究，提出了基于事件触发的分层随机双采样方法 (ESRDS)．通过与传统采样方法的仿真实验比较，证明该方法在分析包长、包到达时间间隔时性能都比较高，尤其该方法采样结果在估计包到达时间间隔时方面和实际比较接近．同时作者还对采样测量方法在计算 Hurst 参数方面进行了研究，应用 ESRDS 对网络流量的 Hurst 参数进行了估计，估计值和实际值之间的误差非常小．

关键词: 采样测量，Hurst 系数，双采样方法

中图分类号: TP393 **文献标识码**: A **文章编号**: 1009-5896(2004)07-1163-05

A Novel Network Traffic Sampling Measurement Method Based on Event-Driven and Its Application Research

Yang Xin-yu Wei Heng-yi Cheng Zhu-lin

(*Dept. of Computer Science and Technology, Xi'an Jiaotong University, Xi'an* 710049, *China*)

Abstract This paper researches mainly on the network traffic sampling measurement methods. The Stratified Random Double Sampling method based on Event-driven(ESRDS) is proposed. It's performance is proved to be better than traditional methods in analyzing the length and interval of packets according to simulation experiments. Especially, estimated interval time of packages by this method is nearly real. It also researches on the measurement methods in the calculating Hurst coefficient. According to the estimation of Hurst coefficient using ESRDS method, the error is much less between estimated and real values.

Key words Sampling measurement, Hurst coefficient, Double Sampling method

1 引言

由于吉比特以太网和其它高速网络技术的发展，对流量分组进行全测量几乎不可能，同时，大量的流量日志也使流量行为分析相当困难，为了解决这一问题，近几年，流量采样测量研究已成为高速网络流量测量的研究重点 [1−4]．

根据采样触发机制的不同，采样测量方法可分为基于事件触发的采样方法和基于时间触发的采样方法 [5]．但这些方法在测量包到达时间间隔这个指标时都存在一些缺点，从而导致误差很大．其主要原因是采样统计本来就是一个近似的方法，而上述方法的样本又不是直接获得的，这就导致了最后误差的扩大．本文提出了基于事件触发的分层随机双采样技术，并应用于单点测量，解决了估计包到达时间间隔时存在的缺点，且在估计网络流量的其他指标上性能也较高．

2 基于事件触发的分层随机双采样方法

基于事件触发的分层随机双采样方法 (Stratified Random Double Sampling method based

[1] 2003-03-12 收到，2003-11-10 改回
国家 863 项目资助课题 (2001AA112111)

on Event-driven, ESRDS) 的基本原理是：对网络流量按包数进行分层，每层中包的个数固定，采样过程在每一层中随机抽取两个相邻的数据包．其示意图如图 1 所示．

图 1 基于事件触发的随机双采样示意图

设每层中数据包的个数为 m ，第 i 层的采样位置为 n_i ，采样位置 n_i 是服从均匀分布的随机数．如网络的全流量 $X=(X_1,X_2,\cdots,X_m,X_{m+1},X_{m+2},\cdots,X_{2m},\cdots,X_{km},X_{km+1},\cdots,X_{(k+1)m},\cdots)$ ，其中 m 为正整数， $k=0,1,2,\cdots$ ，则经 ESRDS 采样后的采集样本为 $S=(X_{n_1},X_{n_1+1},X_{m+n_2},X_{m+n_2+1},\cdots,X_{(k-1)m+n_k},X_{(k-1)m+n_k+1},\cdots)$ ，其中 m 为正整数， $1\le n_k\le m,\ k=0,1,2,\cdots$ ．

ESRDS 有如下几个特点：

(1) 采用事件触发．选用事件触发 (即包计数) 而不采用时间触发，首先是因为时间触发的采样测量方法在测量网络流量的某些参数时性能不高，而且事件触发的方式能够保留网络流量的突发性，这在估计网络流量的自相似性的时候非常重要．

(2) 采用均匀分布的随机数．当网络出现周期性行为的时候，如果采用固定位置采样，那么就有可能只采集到周期行为的一部分．为了避免这种情况，采样位置使用随机数而不选用固定位置．随机采样避免了同步的影响，而且提高了样本随机性，样本的统计量可以得到一个无偏估计．我们采用服从均匀分布的随机数来作为采样位置．

(3) 双采样． ESRDS 采样过程在每次采样的时候都采集两个相邻的数据包，也就是说，首先按照包数对网络流量分层，每层中包的个数固定，采样过程根据生成的随机采样位置，采集该层内第 n 个和第 $n+1$ 个数据包并记录．在对包到达时间间隔进行分析的时候，将这两个相邻数据包之间的到达时间间隔作为数据包到达时间间隔的样本．由于这些样本是真实的包到达时间间隔，而不是经过二次处理得到的，从而提高了估计的准确性．

ESRDS 算法步骤：

(1) 设置每层数据包个数 N ，从第 0 层开始采样；

(2) 随机生成该层采样位置 $n(0\le n<N-1)$ ；

(3) 在该层内连续采集第 n 个和第 $n+1$ 个数据包；

(4) 若流量未结束，则转 (2) ；否则采集结束．

3 ESRDS 与传统采样方法的比较

传统的方法在测量包到达时间间隔的时候都存在比较大的误差 [5] ，在后面的实验中我们主要将事件触发的分层随机双采样和事件触发的系统采样在各个方面的性能上进行比较．

本文以 Berkeley 实验室 1994 年在广域网上监测到的网络流量作为原始数据 (http://ita.ee.lbl.gov/html/contrib/DEC-PKT.html) ．在实验中为了保持相同的采样比率， ESRDS 中的采样间隔是系统采样的两倍．由于需要对两种方法进行比较，所以我们应用了 λ^2 检验法．

3.1 包到达时间间隔

包到达时间间隔是指两个相邻的数据包到达的时间间隔．实验数据为 2000 s 的原始网络流量，约 50 万个数据包．图 2 显示了 ESRDS 和系统采样处理之后的统计结果，其中采样比率为 1/100 ，统计区段为 0.001 s ．从图 2 可以看出， ESRDS 采样结果的包到达时间间隔服从重尾分布，系统采样方法存在很大的误差．

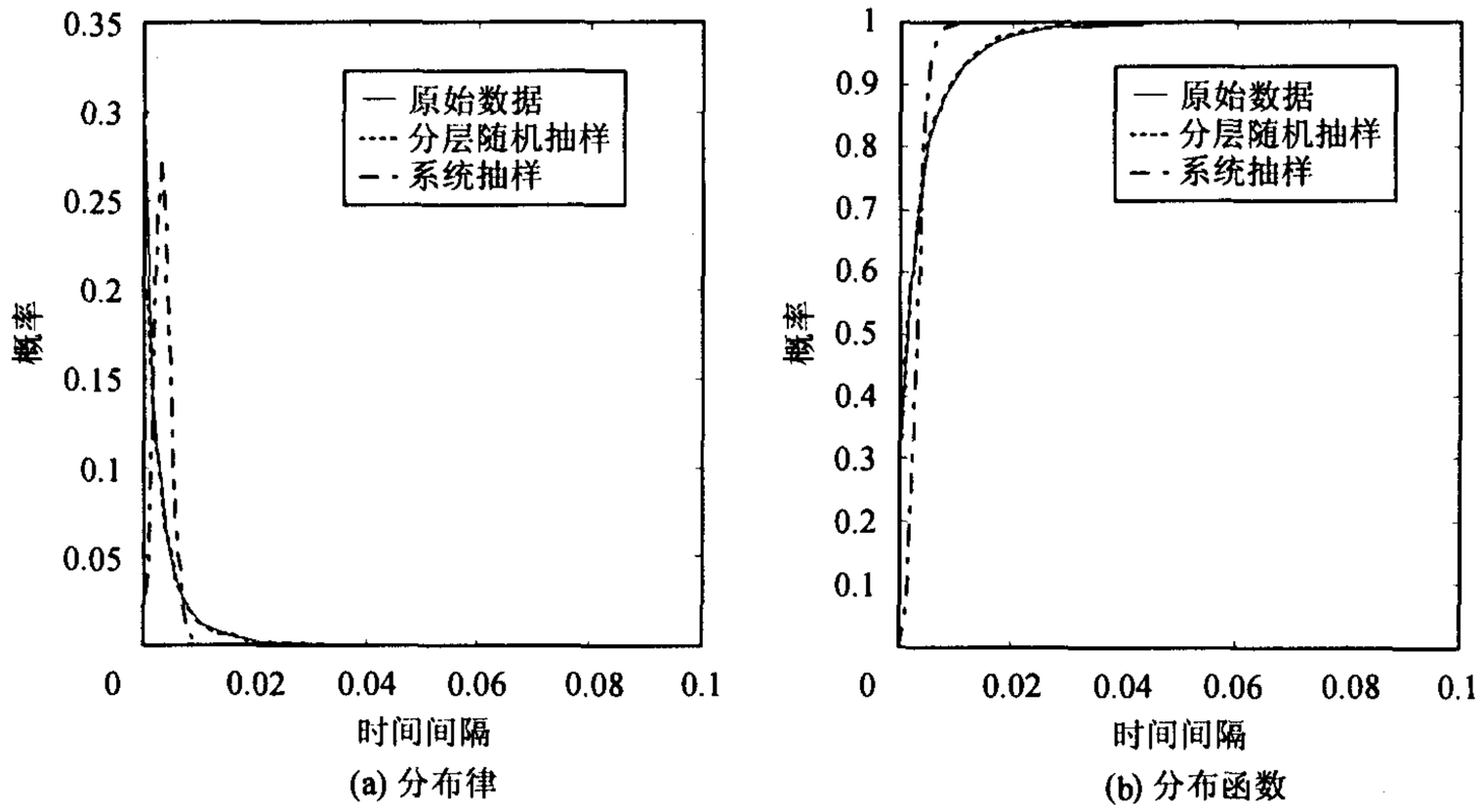

图 2　包到达时间间隔分析实验的分布率与分布函数比较图

表 1 显示了不同情况下的多次实验结果，可见 ESRDS 采样的性能比系统采样要高很多，特别是在采样比率比较大的时候.

3.2 包长

包长也就是每个数据包的字节数. 我们选择 2000 s 的网络流量，约 30 万个数据包. 在不同的采样比率下进行多次实验，实验的分析结果如表 2 所示. 从表 2 可以看出， ESRDS 和系统采样的性能非常接近，且误差都是非常小的. 图 3 是采样比率为 1/100 的情况下，原始数据的包长统计结果和两种采样方法统计结果的比较图.

通过以上的实验分析可知， ESRDS 在分析包长、包到达时间间隔时性能都比较高，尤其是在估计包到达时间间隔时弥补了传统方法存在较大误差的缺陷，所以在单点测量时应用该方法是非常可行的.

表 1　测量包到达时间间隔的多次模拟实验结果 (s)

采样比率	λ^2	
	分层随机双采样 (事件触发)	系统采样 (事件触发)
1/10	0.2545	0.6689
	0.2547	
	0.2565	
	平均： 0.2552	
1/100	0.2817	1.4400
	0.2802	
	0.2810	
	平均： 0.2897	
1/200	0.2902	1.5747
	0.2922	
	0.2952	
	平均： 0.2925	

表 2　测量包长的多次模拟实验结果 (s)

采样比率	λ^2	
	分层随机双采样 (事件触发)	系统采样 (事件触发)
1/10	0.0091	0.0076
	0.0072	
	0.0081	
	平均： 0.0081	
1/100	0.0226	0.0291
	0.0248	
	0.0308	
	平均： 0.0261	
1/200	0.0372	0.0427
	0.0463	
	0.0503	
	平均： 0.0446	

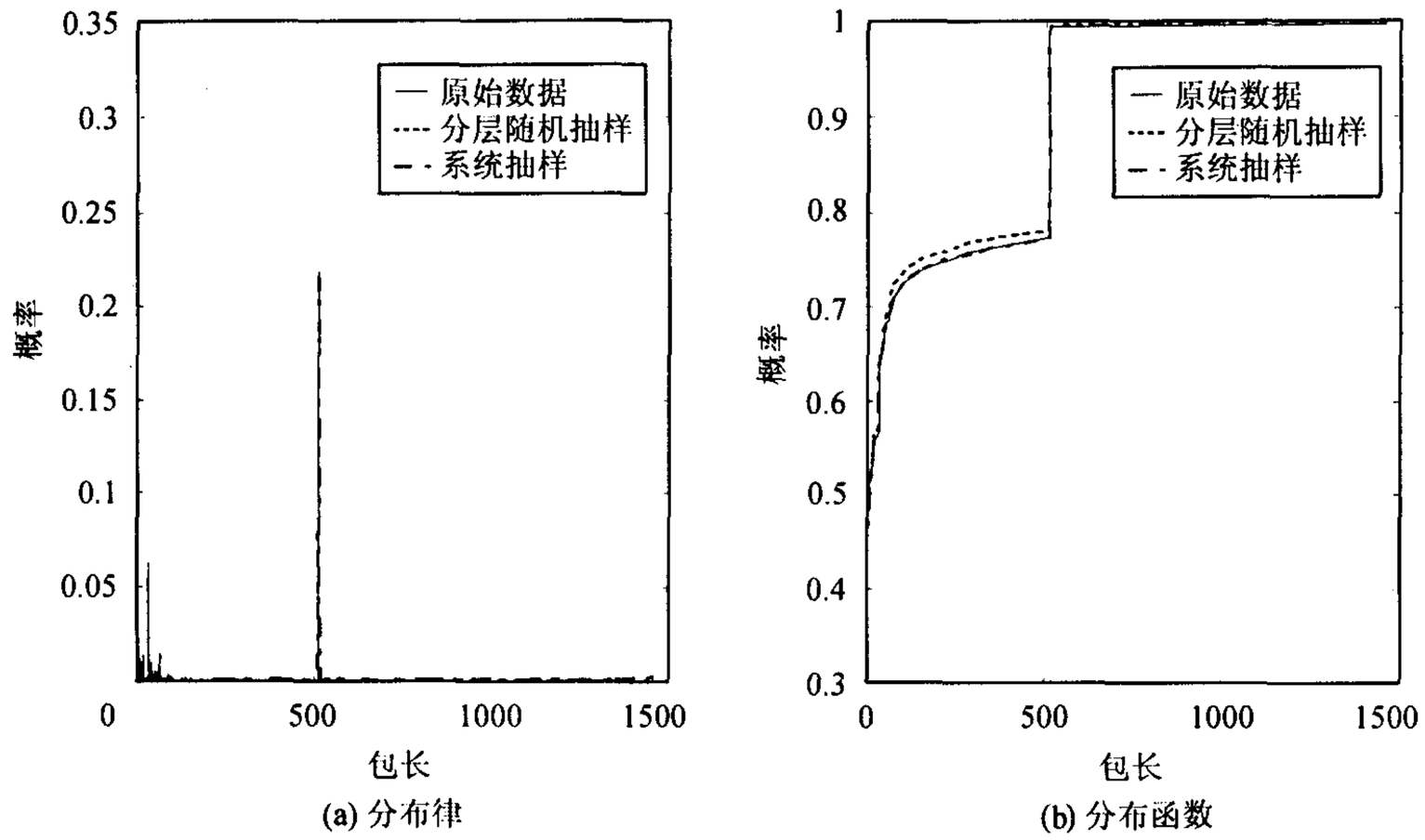

图 3 包长分析实验的分布率与分布函数比较图

4 自相似 Hurst 参数的测量

网络流量具有自相似性，而自相似 Hurst 参数则是自相似性的一个重要度量指标。自相似 Hurst 参数表征了网络流量突发性的程度，这也是我们采用事件触发方式采样的一个重要原因。本文就采样技术在测量自相似 Hurst 参数方面的应用进行了探讨，并应用 ESRDS 方法对网络流量的 Hurst 参数进行了估计。

4.1 估计方法

我们应用时间方差曲线来对网络流量的自相似性进行分析，取 20 个不同的时间单位 ($m = 1, 2, \cdots, 9, 10, 20, \cdots, 90, 100, 200$) 分别计算其样本方差 V_m，通过对 $\ln m$ 和 $\ln V_m$ 做线性拟合，得到 β 的估计值。

在采样的情况下，我们无法获得每个时间单位内的准确包数。故我们根据分层的大小和采样样本来对一个时间单位内的包数进行估计。设进行统计的时间单位为秒，样本的统计结果为：$X = (X_t; t = 0, 1, 2, 3, \cdots)$; X_t 表示第 t 个时间统计区间内采集的样本个数。$Y = (Y_t; t = 0, 1, 2, 3, \cdots)$ 表示第 t 个时间统计区间内实际的包个数。L 表示基于事件触发的分层随机双采样方法中采用的分层大小，采样比率为 $2/L$。那么 Y 的估计值 $\hat{Y}$ 为 $\hat{Y}_i = \begin{cases} [(X_i - 1) \times L]/2, & X_i\text{为奇数} \\ (X_i \times L)/2, & X_i\text{为偶数} \end{cases}$.

4.2 实验分析

图 4 是一次实验结果的 $\ln m$-$\ln V_m$ 的线性拟合图 (原始数据为 2000 s 的实际网络流量)。图 4(a) 是采样比率为 1/100 的情况下，根据采样样本的估算结果 ($H = 0.8614$)，图 4(b) 是原始数据统计结果的拟合图 ($H = 0.8745$)。

5 结论

我们主要对网络流量采样测量进行了研究，针对传统的单点采样测量方法所存在的缺陷，提出了事件触发的分层随机双采样方法。实验证明该方法在分析包长、包到达时间间隔时性能都比较高，尤其是在估计包到达时间间隔时弥补了传统方法存在较大误差的缺陷。同时作者还对采样测量方法在计算 Hurst 参数方面进行了研究，应用 ESRDS 对网络流量的 Hurst 参数进行了估计，估计值和实际值之间的误差非常小。因此，ESRDS 具有较高的准确性，采样精度较高，能够较好地反映网络流量行为。

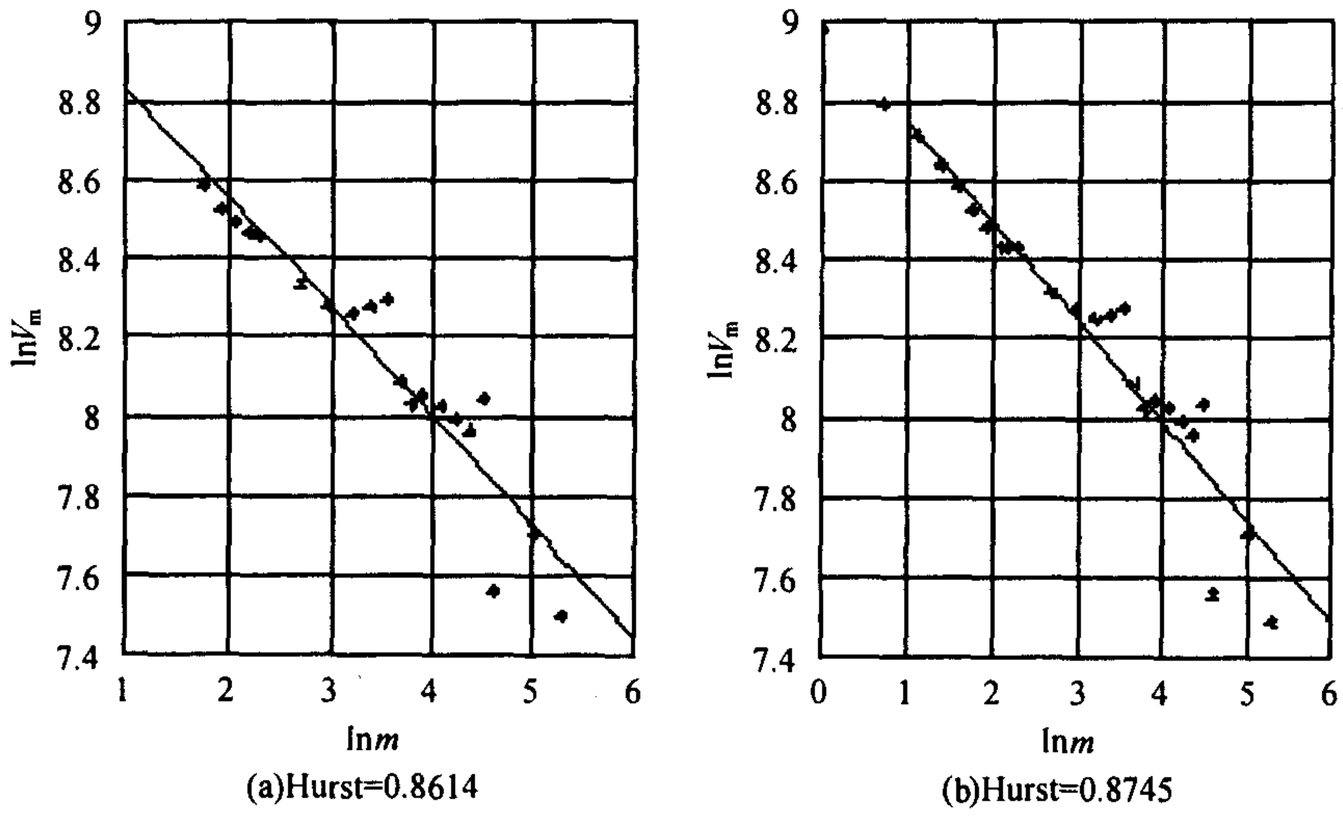

图 4　ln$m$-lnV_m 的线性拟合图

参　考　文　献

[1] Cáceres R, Duffield N G, Feldmann A, *et al.*. Measurement and analysis of IP network usage and behavior. *IEEE Communications Magazine*, 2000, 38(5): 144–151.

[2] Duffeld N G, Grossglauser M. Trajectory sampling for direct traffic observation, *Computer Communication Review*, 2000, 30(4): 271–282.

[3] Estan C, Varghese G. New directions in traffic measurement and accounting. In Proceedings of ACM SIGCOMM, San Francisco, CA, 2002: 323–336.

[4] Duffield N, Lund C, Thorup M. Charging from sampled network usage. ACM SIGCOMM Internet Measurement Workshop 2001, San Francisco, CA, November 1-2, 2001: 101–112.

[5] Claffy K C, Polyzos G C, Braun H W. Application of sampling methodologies to network traffic characterization. *Computer Communication Review*, 1993, 23(4): 194–203.

杨新宇：　男，1973 年生，讲师，博士后，研究方向：计算机网络测量、流量工程与网络行为学.

魏恒义：　男，1963 年生，高级工程师，研究方向：软件集成与开发技术研究.

程竹林：　男，1977 年生，硕士生，研究方向：计算机网络测量.