# 面向传输层协议的网络流量 Hurst 指数分析

朱海婷，丁伟

（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：**描述网络流量自相似性的重要参数 Hurst 指数在近十年以来被广泛接受和使用，但在此期间网络流量成分发生很大变化。本文根据 2005-2010 年间在 CERNET 江苏省网边界采集的多组普通时段和特殊时段的实测数据，使用 R/S 分析方法分别计算总体流量、TCP 流量和 UDP 流量的 Hurst 指数，对其进行分析得出娱乐业务会导致 Hurst 指数上升、在一个连续的时间段内网络流量与 Hurst 指数呈负相关、基于 PPS 计算的 Hurst 指数小于相同条件下基于 BPS 计算的值等结论，并对仿真流量使用的 Hurst 指数数值给出了具体的建议。

**关键词：**网络流量分析；Hurst 指数；自相似

**中图分类号：**TP393　　**文献标识码：**　　**文章编号：**

## Transport Layer Oriented Network Traffic Hurst Exponent Analysis

ZHU Hai-Ting，Ding Wei

(College of Computer Science & Engineering, Southeast University;

Key Laboratory of Computer Network Technology in Jiangsu, Nanjing 210096, China)

**Abstract:** Hurst exponent has been well accepted and widely used as an important indicator of self-similar character in network traffic for the last decade; meanwhile the constitution of network traffic has changed dramatically. Based on real traffic data during 2005 and 2010 collected from CERNET Jiangsu Border router, R/S method is used to analysis Hurst exponent of the sum traffic, TCP traffic and UDP traffic on several regular data and some special time data. Result shows that traffic generated by entertainment business could increase the Hurst exponent, network traffic is negative correlation with Hurst exponent in daily traffic, the Hurst exponent calculated by PPS result is less than the one calculated by BPS result at the same data etc. Recommended Hurst exponent range for simulation is proposed in the end.

**Key words:** network traffic analysis; Hurst exponent; self-similar

## 1 引言

在传统的交换网络中，业务流量到达过程被假定为服从泊松过程，然而当前的 IP 网络，由于采用了数据分组交换传输模式，使得 IP 网络业务流量呈现出与电话交换网络流量截然不同的特性。对各种真实环境中的网络业务流量的测量和分析获得并被广泛承认的结果包括：Leland W.E., Willinger W.等对局域网的研究发现实际以太网网络流量呈现出完全不同于传统泊松模型的特性[1]：重尾特性（Noah 效应）和自相似性（Joseph 效应）；Paxson V.和 Floyd S.对广域网中的 FTP 协议流量进行了测量，分析结果显示 IP 包到达的间隔时间不仅不服从负指数分布，而且不是独立分布的，大部分时候是多个 IP 包连续到达，具有突发性、长程相关性和自相似性[2]；Erramilli A.等人对当时网络中占主导地位的传输协议 TCP 协议进行了分析，讨论了 TCP 协议反馈机制与网络流量自相似性的关系[3]；日本学者 Nakashima 对 TCP 流量的自相似特性的讨论是从资源受限条件下，不同网络参数与自相似性的关系的角度展开的，对各种场景、不同编码的 VBR (Variable Bit Rate) 视频数据进行研究分析也得到了突发性、长相关和自相似性是 VBR 视频流的固有特性的结论[4]。

上述结论要么产生时间较早（参考文献 1-3），当时互联网上绝大部分流量均为 TCP 流量，要么是直接面向 TCP 流量进行的分析（参考文献 4）。通过我们对江苏省网边界路由器到主干线路持续的观测，2005 年的 UDP 流量占总流量 5%左右，而 2009 年这个比例已接近 50%，在这样的情况下，上述结论是否依然存在？本文旨在根据实验

收稿日期：
基金项目：国家 973 计划(No:2009CB320505)；国家科技支撑计划（No:2008BAH37B04）
通信作者：朱海婷(1983－)，女，博士生，从事网络测量研究；E-mail: htzhu@njnet.edu.cn

室利用 WATCHER[5]系统采集、在 IP TASCM[6]平台上保存的 2005 年至今的被动测量数据，通过对整体、TCP 和 UDP 流量分别进行 Hurst 指数计算，来完成以下几个方面的分析：

1 判断自相似现象在新的网络业务和流量下的呈现状态；

2 确定自相似特性与业务流量的变化是否相关；

3 分析使用不同传输协议（TCP 和 UDP）的流量的 Hurst 指数是否有明显区别且与网络流量是否相关；

4 根据实际流量的特性提供给研究者流量发生器的自相似特性 Hurst 指数取值范围。

# 2 数据和算法

## 2.1 实验数据

本文试验数据来源于 CERNET 江苏省网边界到主干的线路上分光后对双向流量报文头部的被动采集和保存，起始于 2005 年 11 月，以固定长度采集所有报文。由于存储资源限制和线路改造升级的原因，2007 年下半年以后改为 1/4 流抽样的方式进行采集和存储。2009 年 12 月前每次采样持续时间为 1 小时，在此之后由于硬件平台的改善，增加了数次连续 24 小时的采样。

为了便于对照，本实验方案选取的分析数据如表 1 所示，分为三个对比组，7 组数据。第一组数据为“远期数据”，采于 2007 年 1 月 6 日前，均为 1 小时数据。第二组数据采于 2010-4-21 日，当日为青海玉树地震悼念日，为 2 小时 1/4 流抽样的采集数据。第三组是为“近期数据”，是 24 小时 1/4 流抽样采集数据。

表 1 流量数据列表

| 编号 | 数据名称 | 采集时间 | 数据量 | TCP 流量比例 % | | UDP 流量比例 % | |
|---|---|---|---|---|---|---|---|
| | | | | 入流量 | 出流量 | 入流量 | 出流量 |
| 1-1 | 2005-11-10 | 14:00-15:00 | 149G | 94.76 | 94.7 | 4.91 | 5.04 |
| 1-2 | 2007-1-6 | 14:00-15:00 | 150G | 78.39 | 77.35 | 21.75 | 22.55 |
| *2* | *2010-4-21* | *16:00-18:00* | *125G* | *43.59* | *52.63* | *56.32* | *46.59* |
| 3-1 | 2009-12-17 | 00:00-24:00 | 856G | 56.27 | 71.11 | 43.37 | 28.53 |
| 3-2 | 2010-1-12 | 00:00-24:00 | 942G | 45.01 | 60.98 | 53.98 | 38.40 |
| 3-3 | 2010-5-18 | 00:00-24:00 | 976G | 54.51 | 72.86 | 45.17 | 26.45 |
| 3-4 | 2010-6-15 | 00:00-24:00 | 844G | 49.30 | 60.34 | 50.57 | 38.32 |

## 2.2 Hurst 指数定义及计算方法

Hurst 指数是反映网络流量长相关和自相似特性的重要指标并且可以作为业务突发的度量。Hurst 指数在网络流量中其值落在[0.5,1]，Hurst 指数的值越大表明自相似程度越高。

本文采用 R/S 分析(Re-scale Range Analysis)又称为重标极差分析法。Hurst 指数反映了时间序列均值的累计离差随时间产生的变化范围，并且建立了以下关系 $\left[R/S\right]_k = Ck^H$ (1)

其中 R 为时间序列的极差，S 为标准差(即重标极差)，k 为时间间隔，C 为常数，H 为 Hurst 指数。

假设数据为 $X=\{x_t\}_{t=1}^{N}$，在确保数据为平稳随机过程的前提下。算法为：

1. 初始条件为 N 个数据，数据子序列个数 m=1 组;
2. 如果数据的序列的长度 k=N/m 大于阈值，进入步骤 3；否则进入步骤 4；
3. 计算 m 个子序列的均值

$$E[(x)_k]=\frac{1}{k}\sum_{i=1}^{k}x_i \quad (2)$$

累计偏差

$$X(i,k)=\sum_{j=1}^{i}[x_j-E[(x)_k]]=\sum_{j=1}^{i}x_j-i*E[(x)_k],1\le i\le k \quad (3)$$

极差 $R(k)=\max X(i,k)-\min X(i,k),1\le i\le k$ (4)

标准差 $S(k)=\left\{\frac{1}{k}\sum_{i=1}^{k}[x_i-E[(x)_k]^2\right\}^{\frac{1}{2}}$ (5)

记录 R/ S 均值和对应子序列长度 k，m 加 1 后返

回步骤 2;

4. 通过记录的极差与标准差的比值 R/ S，根据公式(1)变换

$$\ln\frac{R(k)}{S(k)} = \ln(C) + H * \ln(k)\ ,\quad 0 \le H \le 1 \qquad (6)$$

通过最小二乘法回归计算出 Hurst 指数 H，算法结束。

### 3.1 面向总体的比对

采用上述算法，将分析对象分为：入流量/出流量(Bps)，入报文/出报文（pps）,并面向总流量、TCP 流量和 UDP 流量分别计算。实际计算中为保证数据的稳定性，R/S 方法中的 k 的阈值取值为 10，并将数据 2 根据其持续时间等分成 2-1 和 2-2。最终获得的样本总数为 12 类，96 个见表 2。E1、E2、E3 分别是三组结果对应的均值。

## 3 结果与分析

表 2 各组数据流量 Hurst 指数

| **BPS** | SUM in | TCP in | UDP in | SUM out | TCP out | UDP out | **PPS** | SUM in | TCP in | UDP in | SUM out | TCP out | UDP out |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1-1 | 0.83 | 0.83 | 0.95 | 0.82 | 0.83 | 0.95 | 1-1 | 0.82 | 0.82 | 0.94 | 0.82 | 0.81 | 0.95 |
| 1-2 | 0.90 | 0.91 | 0.89 | 0.86 | 0.87 | 0.84 | 1-3 | 0.87 | 0.90 | 0.86 | 0.86 | 0.87 | 0.85 |
| **E1** | **0.87** | **0.87** | **0.92** | **0.84** | **0.85** | **0.90** | **E1** | **0.85** | **0.86** | **0.90** | **0.84** | **0.84** | **0.90** |
| *2-1* | *0.72* | *0.79* | *0.75* | *0.76* | *0.77* | *0.86* | *2-1* | *0.70* | *0.71* | *0.72* | *0.67* | *0.71* | *0.77* |
| *2-2* | *0.74* | *0.80* | *0.80* | *0.80* | *0.80* | *0.90* | *2-2* | *0.71* | *0.74* | *0.76* | *0.73* | *0.75* | *0.81* |
| **E2** | **0.73** | **0.80** | **0.78** | **0.78** | **0.79** | **0.88** | **E2** | **0.71** | **0.73** | **0.74** | **0.70** | **0.73** | **0.79** |
| 3-1 | 0.84 | 0.85 | 0.87 | 0.86 | 0.87 | 0.89 | 3-1 | 0.81 | 0.82 | 0.84 | 0.82 | 0.83 | 0.85 |
| 3-2 | 0.82 | 0.83 | 0.86 | 0.82 | 0.83 | 0.86 | 3-2 | 0.80 | 0.79 | 0.83 | 0.78 | 0.80 | 0.83 |
| 3-3 | 0.82 | 0.83 | 0.86 | 0.80 | 0.81 | 0.87 | 3-3 | 0.79 | 0.80 | 0.82 | 0.77 | 0.78 | 0.83 |
| 3-4 | 0.87 | 0.89 | 0.90 | 0.87 | 0.89 | 0.89 | 3-4 | 0.83 | 0.84 | 0.85 | 0.82 | 0.83 | 0.85 |
| **E3** | **0.84** | **0.85** | **0.87** | **0.84** | **0.85** | **0.88** | **E3** | **0.81** | **0.81** | **0.84** | **0.80** | **0.81** | **0.84** |

下面从不同角度对这些样本和均值进行对比和分析：

- 通过第一组和第三组数据的均值对比可以看出得出该信道 Hurst 指数整体略有下降，但仍旧处在较高的水平。这说明在网络流量增大的条件下，网络的自相似特性对传输层协议的使用导致的流量成分的变化不敏感。
- 第二组数据的 Hurst 指数明显小于另外 2 组，由于当天是玉树地震哀悼日，因此可以认为是娱乐业务流量减少的原因导致，这说明该类流量是影响网络 Hurst 指数的重要原因，所以自相似特性与业务流量种类有着较为密切的关系。
- 对比同一数据按 BPS 和 PPS 的计算结果。一共 48 对可比对数据，除 1-2 数据的 UDP/out 外，其余占总数 98%的 47 对数据 Hurst 值按 BPS 计算均不小于按 PPS 计算。因此，可以认为按 BPS 计算的 Hurst 指数略大于按 PPS 计算。
- TCP 流量与 UDP 流量的 Hurst 指数比对。一共有 32 组可比对数据，其中 27 组 UDP 的 Hurst 指数不小于 TCP 的 Hurst 指数，占 84.4%。从均值角度分析该项特性，12 组可比对数据，11 组 UDP 的值大于 TCP 的值，差值范围在[-0.02，0.09]之间，据此可以认为 UDP 流量的 Hurst 指数略大于 TCP 流量的 Hurst 指数。
- 总体 Hurst 指数与对应数据的 TCP/UDP 流量 Hurst 指数关系。从均值数据对比可以看出按总流量计算的 Hurst 指数要小于分类计算的结果。

根据上面最后三项分析的结果，我们建议当使用 Hurst 参数生成模拟流量时，如果使用面向总体流量模型，Hurst 指数均值可取 0.83。如果采用更精确的面向传输层以上的流量模型则建议 TCP 流量采用 Hurst 指数均值为 0.85，对应的 UDP 流量为 0.88。如果采用 PPS 产生流量，则上述各数值还应对应减少 0.03.

### 3.2 连续时间段的 Hurst 指数变化分析

由于第三组 4 条数据均为连续 24 小时的 Trace 数据，本小节将每条数据以小时为单位分成 24 份，然后分别计算 in/out、TCP、UDP 和总体的 Hurst 指数，并将其与对应时间段内的总流量进行对比。结果如图 1 所示。该图的数据源为 3-1，第三组数

据中其余的3组我们也进行了计算，结果是相似的。从该图可以看出面向总体流量、TCP 流量、UDP 流量的 Hurst 指数在一天内在( 0.69，0.96 )范围内波动，且与对应的流量呈现负相关关系。特别需要指出的是，从该图还可以明显看出在 Hurst 指数的拐点位置也对应着流量的拐点。

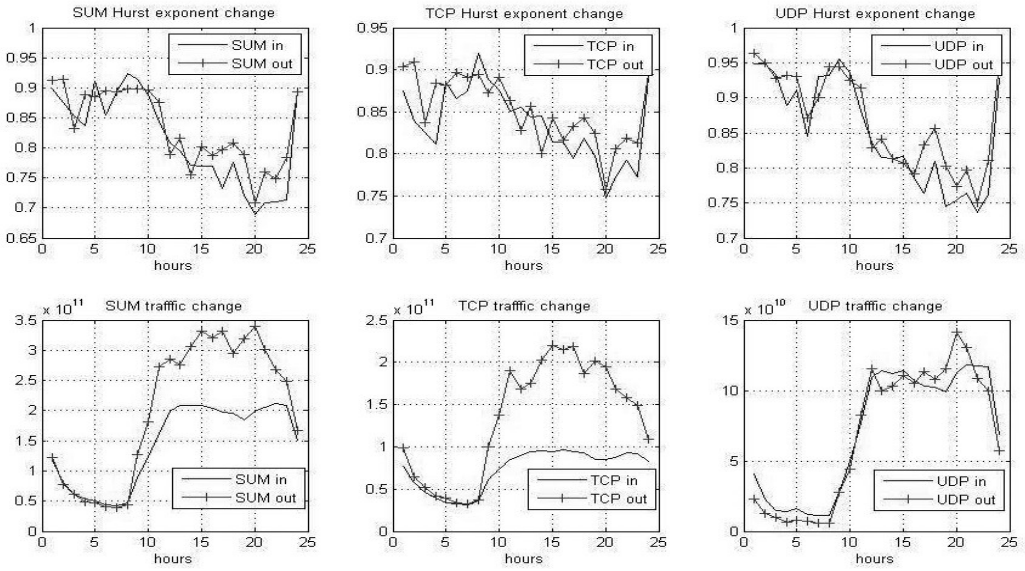


图 1　24 小时中三种 Hurst 指数变化与对应流量变化关系

根据上述分析，我们进一步建议用于模拟流量生成的 Hurst 指数，如果能随流量的大小而变化将更加符合实际情况。

## 4 总结

随着时间的推移，网络业务的变化导致网络流量中的某些特性随之发生变化，这些变化对网络管理和网络工程有很重要的意义。在观察到流量成分发生巨大变化的情况下，本文根据观察到的 CERNET 江苏省网边界流量，从总体、TCP 和 UDP 三个角度分析了流量的自相似特性。得到的结论包括：Hurst 系数整体呈轻微下降趋势、TCP 流量的 Hurst 指数低于 UDP，但高于总体流量、按 BPS 计算的 Hurst 指数高于按 PPS 计算和 Hurst 指数与流量呈负相关关系等，并对模拟流量生成时的 Hurst 指数取值给出了具体的建议。

由于时间的原因，本文的结论均来自于直观的比对，因此也还停留在比较简单的现象描述的层面上。今后相关的研究工作将首先深入分析这些现象的成因，在此基础上选择适合的数学工具，从定量角度对相关结果进行更加规范的描述。

## 5 参考文献

[1] Leland WE, Taqqu MS, Willinger W, et al. On the Self-similar Nature of Ethernet traffic (Extended Version) .IEEE/ACM Transactions on Networking, 1994, 2 (1):1~15.

[2] Paxson V., Floyd S. Wide area traffic: the failure of Poisson modeling. IEEE/ACM Transactions on Networking, 1995, 3(3):226~244.

[3] Erramilli A., Roughan M., Veitch D., et al. Self-similar traffic and network dynamics. Proceedings of the IEEE, 2002, 90(5):800~819.

[4] Nakashima T., Sueyoshi T. Self-Similar Property for TCP Traffic under the Bottleneck Restrainment.in:21st International Conference on Advanced Information Networking and Applications Workshops, 2007(AINAW'07).2007.228~233.

[5] 程光，丁伟，龚俭．高速网络流量通用测量平台-WATCHER [C]．见: CSIT 2004 会议论文集, 南京, 2004．122-128．

[6] 朱海婷，丁伟，夏震等。开放式网络测量数据分析系统 IP TASCM[C].见: CERNET 2009 会议论文集，天津，2009