2006－DPS－126（33）
2006－CSEC－32（33）
2006／3／17

シミュレーションによるHeavy-tailedトラフィックの特性解析

中嶋 卓雄, 築地原 護

九州東海大学 応用情報学部
熊本市渡鹿9丁目-1-1
E-mail: {taku@ktmail,40mie102@stmail}.ktokai-u.ac.jp

スケール不変なバースト性や自己相似性は実際のネットワークで発見されており，この自己相似性とネットワークやシステムパラメータとの関係は主にend-to-endのデータ転送の環境において議論されてきた．この自己相似性は主にWebサーバのファイルサイズのHeavy-tailedな分布やユーザセッションの間隔が原因すると言われている．しかし，Webサーバへのアクセスは均一な分布ではなく，特定のサーバに収集する傾向があり，また他のネットワークパラメータの要素がどのように自己相似性と関係あるのかについて厳密には語られていない．本研究では，ネットワークシミュレータを利用してネットワーク環境を変え自己相似なトラフィックの特性を抽出した．シミュレーションの結果から，次のような結論を得ることができた．まず，第一にファイルサイズのHeavy-tailedな分布が，特に小さなαの値に対して，自己相似性を導出する．第二に，パワー則に従ったデータレートの分布は自己相似性を強調する．最後に，より大きなエラーレートはスループットの振幅を活発かさせ，自己相似性を維持する．

Property Analysis of Heavy-tailed Traffic by Simulator

Takuo Nakashima, Mamoru Tsuichihara

Department of Information Science Kyushu Tokai University
9-1-1 Toroku, Kumamoto, Japan
E-mail: {taku@ktmail,40mie102@stmail}.ktokai-u.ac.jp

The scale-invariant burstiness or self-similarity has been found in real network. Relation between self-similarity and network and/or system parameter is mainly discussed in the context of end-to-end data transmission environment, and this self-similarity is mainly caused by the file size of Web servers or the duration of user sessions On the other hand, the accesses to Web servers are not uniformly distributed, but are confined to specific Web servers such as search engine sites or portal sites, and other elements of network parameters are not clearly described in the context of self-similarity. In this paper, we have investigated the property of self-similar traffic varying the network environment using the network simulator. After analyzing the simulated results, following properties were extracted. Firstly, heavy-tailed distribution of file size induces self-similarity, especially in the case of small α. Secondly, distribution with power law of data rate emphasize self-similarity. Finally, greater error rate induces to activate the fluctuation of the throughput, and remains self-similar property.

1 Introduction

Since the seminal study of Leland, et al. [1], the scale-invariant burstiness or self-similarity has been found in real network. Leland, et al. [1] demonstrated self-similarity in a LAN environment, Paxson et al. [3] showed self-similar burstiness in pre-WWW WAN IP traffic, and Crovella et al. [2] showed self-similarity for WWW traffic. Currently, self-similarity of network traffic has been widely adopted in the modeling and analysis of network performance.

Relation between self-similarity and network and/or system parameter is mainly discussed in the context of end-to-end data transmission environment. Crovella et al. [2] indicate that this self-similarity is mainly caused by the file size of Web servers or the duration of user sessions, and ftp traffic has the heavy-tailed property of Pareto distribution with $0.9 \leq \alpha \leq 1.1$[3]. Park et al. [4] examined simulations with diverse conditions using network simulator, and discussed the meaning of effect of each element of simulating conditions.

If the access to Web servers are uniformly distributed to each Web server, then Internet traffic are composed in same distribution. The access, however, tend to concentrate to search engine sites or portal sites. The healthy Internet maintain the low error rate under one percentage. But some edge nodes of small provider temporally rise the error rate under the heavy load condition. These error-prone traffics are controlled by the flow control mechanism of TCP. It is the ongoing topic to investigate how this feedback mechanism works the property of Internet traffic.

In this paper, we investigate the property of self-similar traffic varying the network environment using the network simulator (ns-2) [5]. First, the concentrated phenomena for the access to Web servers are to be simulated followed by consideration of the error-prone condition with 10% error rate. Lastly, the discussion of data multiplexing will be presented. We have referred to research by Park et al. [4] in related sections.

This paper is organized as follows. First, the background of heavy-tailed property and Hurst parameter is described in Section 2 followed by the environment of network model and simulation in Section 3. With this simulation, we discuss the effect of self-similar traffic varying the network environment in Section 4. Section 5 is the summary and discussion for future work.

2 Background

First, a definition of heavy-tailed distribution will be given, followed by the relation between shape parameter of Pareto distribution and Hurst parameter.

2.1 Heavy-tailed Distribution

A random variable X has a heavy-tailed distribution if

$$P[X > x] \sim cx^{-\alpha}, \quad 0 < \alpha < 2, \tag{1}$$

for some positive constant c, where $a(x) \sim b(x)$ means $\lim_{x\to\infty} a(x)/b(x) = 1$, and α is called the tail index or shape parameter. Regardless of the behavior of the distribution for small values of the random variable, if the asymptotic shape of the distribution is power law, it is heavy-tailed. If $P[X > x]$ is heavy tailed then X shows very high variability. X has infinite variance, and, if $\alpha \leq 1$, X has infinite mean.

The simplest heavy-tailed distribution is the Pareto distribution. The Pareto distribution is power law over its entire range; its probability density function is

$$p(x) = \alpha k^{\alpha} x^{-\alpha-1}, \quad \alpha, k > 0, \quad x \geq k, \tag{2}$$

and its cumulative distribution function is given as

$$F(x) = P[X \leq x] = 1 - (k/x)^{\alpha}. \tag{3}$$

The parameter k represents the smallest possible value of the random variable, and is called the location parameter.

To assess the property of heavy-tailedness, we employ $log - log$ plots of the complementary cumulative distribution given as

$$\overline{F}(x) = 1 - F(x) = P[X > x] = (k/x)^{\alpha} \tag{4}$$

for the random variable X. Plotted in this way, heavy-tailed distributions have the property that

$$\frac{d \log \overline{F}(x)}{d \log x} = -\alpha, \quad x > \theta \tag{5}$$

for some θ. To check the property of heavy-tailedness, we form $\log - \log$ plots, and look for approximately linear behavior with slope $-\alpha$ in the tail.

2.2 Hurst Parameter

A quantitative measure of self-similarity is obtained by using the Hurst parameter H, which expresses the speed of decay of a time series' autocorrelation function. A time series with long-range dependence has an autocorrelation function of the form

$$r(k) \sim k^{-\beta} \quad as \;\; k \to \infty,$$

where $0 < \beta < 1$. The Hurst parameter is related to β via

$$H = 1 - \frac{\beta}{2}.$$

Heavy-tailed property causes the long-range dependence that is one of the properties of self-similarity. Park[6] presented that heavy tails lead to predictability, and in relation, they lead to long-range dependence in network traffic. Hurst parameter is related to the tail index by $H = (3-\alpha)/2$, which can be predicted by the on/off model in an idealized case corresponding to a fractal Gaussian noise process.

3 Network Model and Simulation

3.1 Network Model

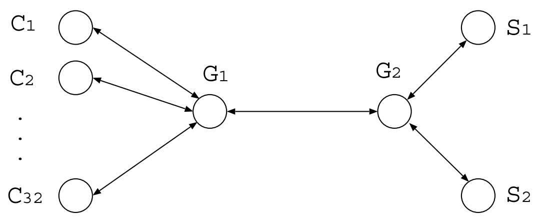


Figure 1. Network Configuration

The network is constructed by nodes and links, both having a buffer, bandwidth, and latency. Server node $s_i(i = 1, \cdots, n)$ have a probability density function $p_i(X)$, where $X \geq 0$ is a random variable denoting file size. In our previous work [7], we measured file size of top page on diverse Web servers and observed heavy-tailed distribution of file sizes. We assume that these actual file sizes are mapped to the probability density function $p_i(X)$, which is called as file size distribution in this paper.

Figure 1 shows 2-server, 32-client network configuration. The link between gateway G_1 and G_2 is shaped as a bottleneck link. We will refer to the traffic from G_2 to G_1 as downstream traffic and traffic from G_1 to G_2 as upstream traffic. Downstream traffic denotes multiplexed file transmission from servers.

3.2 Simulation Setup

We used the LBNL network simulator (ns) [5] to evaluate the effect of different network parameters. The ns is a very popular software for simulating advanced TCP/IP algorithms and protocols. In this paper, we focus on the file size distribution in conjunction with data rate and effect of different flow control algorithms, different TCP modules, with varying error rate.

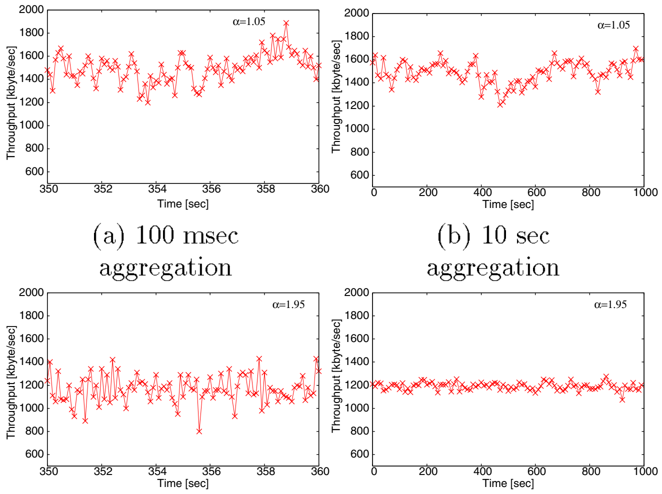
The topology of our simulation is limited to the 2-server 32-client bottleneck configuration in Figure 1. The bottleneck link, however, was varied from 1.5 Mbps to 10 Mbps. Non-bottleneck links were set at 10 Mbps and the latency of each link was set to 15 ms presenting domestic links. The maximum segment size was fixed at 1 kb. We measured the downstream traffic from G_2 to G_1.

4 Results of Simulation

4.1 Self-similar Traffic

Firstly, we examine that transfer of files with Pareto distribution, i.e. heavy-tailed distribution, generate self-similar traffic. The run of this experiment was executed for 1000 simulated seconds with fixed data rate (200kbps) and 10 Mbps bottleneck link. Data source was limited only to S_1 from which 32 data connections were established to 32 clients with TCP/Reno agent. Figure 2 and 3 show that the throughput of traffics generated by Pareto and Exponential distribution are fluctuating over all time series independently. The process with Pareto distribution reveals the self-similar process, which has the heavy-tailed property, and the random process has the exponential distribution. The throughput ranges of both Figures is 500 to 2000 kbps.

Two aggregation levels are plotted in both Figures, (a),(c) in Figure 2 and (a) in Figure 3 - all of three to exhibit 100 ms aggregation time unit - and (b),(d) in Figure 2 and (b) in Figure 3, all having 10 s time unit. Two different location parameter α values, $\alpha = 1.05$ and $\alpha = 1.95$, are shown in (a),(b) and (c),(d) in Figure 2 respectively. In the case that α is close to 2, which is (d), fluctuation changes smoothly at large time scales. It indicates that weak dependency structure appears in time series.



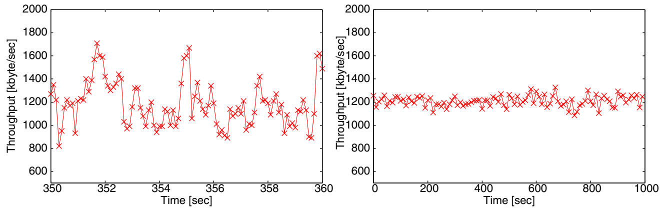
(a) 100 msec aggregation

(b) 10 sec aggregation

(c) 100msec aggregation

(d) 10 sec aggreagtion

Figure 2. Pareto Distribution of File Size



(a) 100 msec aggregation

(b) 10 sec aggregation

Figure 3. Exponential Distribution of File Size

On the other hand, in the case that α is close to 1, which is (b), burstiness is still preserved at large time scales, indicating that self-similarity is observed at α closing to 1. Figure 3 illustrates the throughput generated by Exponential file distribution. We observed that the aggregation time series between exponential and Pareto with $\alpha = 1.95$ are qualitatively indistinguishable.

Examinations were carried out under the condition that the bottleneck link is constrained for the bandwidth with 1.5 Mbps. Park et al. [4] illustrated that resource limitations led to no significant variations for self-similarity. On the other hand, the results of $\alpha = 1.05$ in our experiments, which is not shown, illustrates that the time scale-invariant property disappeared and converged to flat fluctuation in 10 seconds aggregation as well as the case of exponential distribution. This results indicates that bandwidth-constrained bottleneck link reduces the property of self-similarity. Our simulations under a few conditions showed some difference from those reported by Park et al. [4]. We should confirm the effect of differency. Park et al. also reported that selfsimilar property decay in proportion to the strictness of limitation.

4.2 Effect of Data Rate

The experiment described previously was examined with fixed data rate. The accesses to Web servers, however, are not uniformly distributed, but are confined to specific Web servers such as search engine sites or portal sites. We represent this mal-distribution of access to Web server as the data rate distribution with power law. Figure 4 shows the data distribution for each traffic.

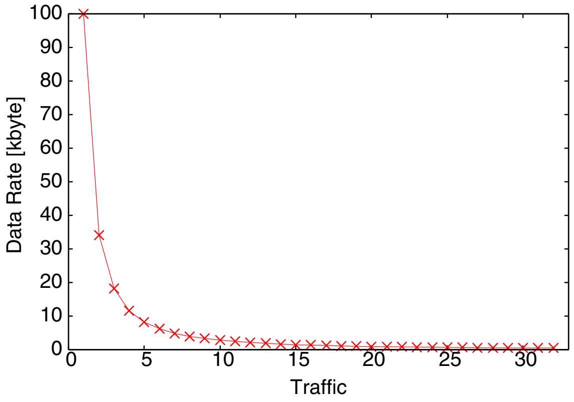
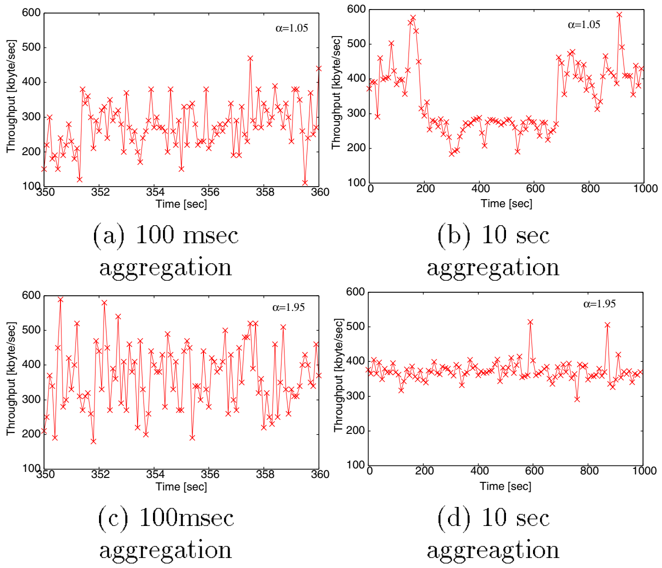


Figure 4. Data Distribution for Each Traffic

Figure 5 illustrates the effect of variant data rate. Two aggregation levels, 100 milliseconds and 10 seconds, and two different α values, $\alpha = 1.05$ and $\alpha = 1.95$, are shown in this Figure 5. Direct comparison of the throughput between Figure 2 and 5 is not appropriate due to the different data rate, but with a different scale, mal-distribution of data rate emphasize self-similarity in small $\alpha(= 1.05)$. In the real network, if the file size distribution has small α, then the self-similarity will be emphasized.

4.3 Effect of Error Rate

In this experiment, we changed the error rate to 10 % on the bottleneck link and examined in 10000 simulated seconds. Figre 6 illustrates the 100 milliseconds and 10 seconds aggregation to compare, and the data of 100 seconds aggregation are shown in Figure 7. Figure 6 and 7 are illustrated in same horizontal scale of Figure 2.



(a) 100 msec aggregation

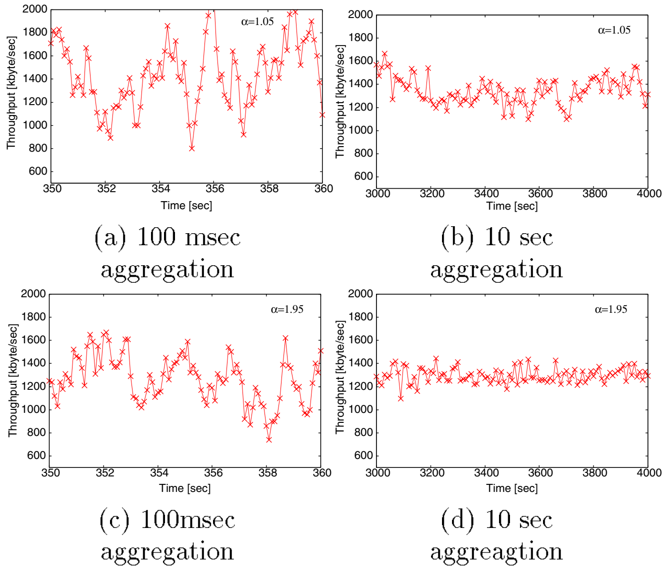
(b) 10 sec aggregation

(c) 100msec aggregation

(d) 10 sec aggreagtion

Figure 5. Effect of Data Rate

These figures indicate that the higher the error rate increases, the larger the range of fluctuation becomes to make fluctuation eventually converged to the same level at coarser aggregation, i.e. 100 seconds. These occurrence of error activate the flow-control mechanism, then the throughput fluctuates in larger range. The self-similarity remains still despite occurrence of the error. On the other hand, UDP traffics with same error rate remain the the same fluctuations under the condition of no error rate.



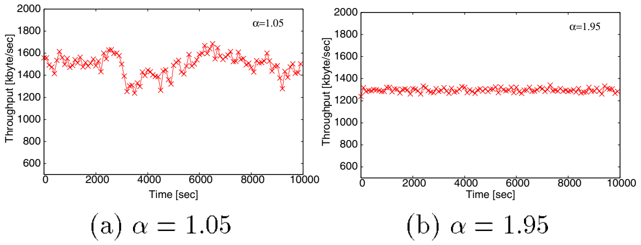
(a) 100 msec aggregation

(b) 10 sec aggregation

(c) 100msec aggregation

(d) 10 sec aggreagtion

Figure 6. Effect of Error Rate with 10%

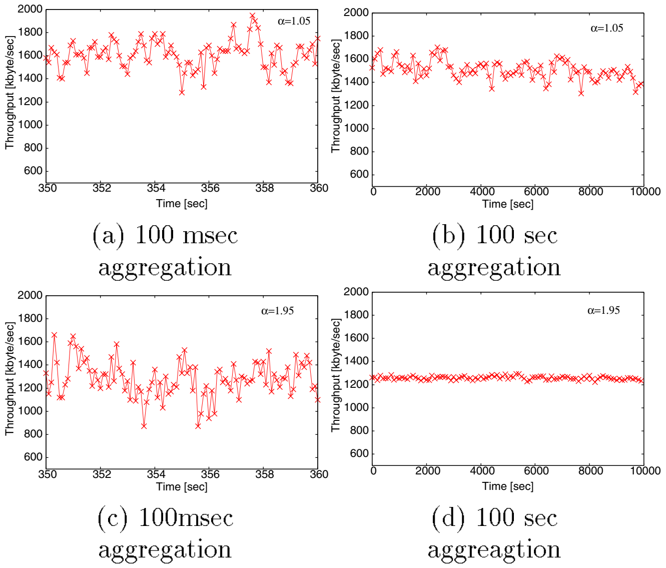


(a) $\alpha = 1.05$

(b) $\alpha = 1.95$

Figure 7. 100 Seconds Aggregation of Error Rate with 10%

Figure 8 shows the effects of 5% error rate, and the two aggregation levels, i.e. 100 milliseconds and 100 seconds. (a),(b) and (c),(d) in Figure 8 illustrate the value of $\alpha = 1.05$ and $\alpha = 1.95$ respectively. Compared to Figure 2 and 6, this Figure illustrates the middle level fluctuations in both Figure 2 and 6. This indicates that error rates steadily affect the all time scale invariance and larger error rate generates larger fluctuation level for the throughput.



(a) 100 msec aggregation

(b) 100 sec aggregation

(c) 100msec aggregation

(d) 100 sec aggreagtion

Figure 8. Effect of Error Rate with 5%

4.4 Effect of Other Elements

Firstly, we examined the effect of traffic multiplexing to generate the new traffic from server S_2 to 32 clients with the same fixed data rate (200 kbps). The result indicated no clear difference between the one traffic flow from S_1 and two traffic flows from S_1 and S_2. This means the property of self-similarity remain existed in terms of traffic multiplexing.

Second experiment is to check how a type of TCP agent effects differently to self-similarity. Then we examined the agent of Tahoe, NewReno and Vegas with the same error rate (10%). Results show that every TCP agent worked the same way, and did not clearly affect the fluctuation. This also means that different flow-control mechanisms maintain the property of self-similarity.

5 Conclusion

In this paper, we have investigated the property of self-similar traffic varying the network environment using the network simulator. After analyzing the simulated results, following properties were extracted. Firstly, heavy-tailed distribution of file size induces self-similarity, especially in the case of small α. On the other hand, constrained bottleneck link reduced the time scale-invariant property. Secondly, distribution with power law of data rate emphasize self-similarity. When the access traffics were tend to distribute under the power law, self-similarity, were emphasized. Thirdly, greater error rate induces to activate the fluctuation of the throughput, and remains self-similar property. The fluctuation level of 10 % error rate was similar to the fluctuation at ten times aggregation level with no error rate. Finally, the effect of traffic multiplexing with two servers, and different TCP agents did not clearly affect the self-similar property.

These experiments were evaluated by the fluctuation pattern on different time scale aggregations. To evaluate the property of self-similarity qualitatively, estimation of Hurst parameter is required. In the next step, we will conduct to estimate Hurst parameter using variance-time plot or R/S estimation.

In the future, the real network experiment is to be conducted followed by the discussion of the precision of the network simulator to find the new metrics for the self-similarity.

References

[1] Leland, W., Taqqu, M., Willinger, W. and Wilson, D.: On the Self-Similar Nature of Ethernet Traffic (Extended Version), *IEEE/ACM Trans. on Networking*, Vol.2, No.1, pp.1-15, February (1994).

[2] Crovella, M and Bestavros, A: Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes, *IEEE/ACM Trans. on Networking*, Vol.5, No.6, pp.835-845 (1997).

[3] Paxson, V. and Floyd, S.: Wide-Area traffic: the failure of Poisson modeling, *IEEE/ACM Trans. on Networking*, Vol.3, No.3, pp.226-244 (1995).

[4] Park, K. ,Kim, G. and Crovella, M. E.: The Protocol Stack and Its Modulating Effect on Self-similar Traffic, *Self-Similar Network Traffic and Performance Evaluation*, K.Park and W.Willinger,Eds.,Wiley-Interscience, New York, pp.349-366, (2000).

[5] UCB/LBNL/VINT groups. UCB/LBNL/VINT Network Simulator, http://www.isi.edu/nsnam/ns/, May (2001).

[6] Park, K. ,Willinger, W.: Self-similar Network Traffic: An Overview *Self-Similar Network Traffic and Performance Evaluation*, K.Park and W.Willinger,Eds.,Wiley-Interscience, New York, pp.1-38, (2000).

[7] Nakashima, T.,Tsuichihara, M.: Stability in Web Server Performance with Heavy-Tailed Distribution, *Proc. of 9th International Conference of Knowledge-Based Intelligent Information and Engineering Systerms(KES 2005)*, Vol. 1, pp.575-581, (2005).