Proceedings of 12th IEEE International Conference on Computer Communications and Networks (ICCCN), 2003

Effect of TCP on Self-Similarity of Network Traffic

Nawaporn Wisitpongphan, Jon M. Peha
Carnegie Mellon University
Pittsburgh, PA 15213
E-mail: nawaporn@andrew.cmu.edu, peha@cmu.edu

Abstract - It is now well known that Internet traffic exhibits self-similarity, which cannot be described by traditional Markovian models such as the Poisson process. In this work, we simulate a simple network with a full implementation of TCP-Reno. We also assume Poisson arrivals at the application layer specifically to determine whether TCP can cause self-similarity even when input traffic does not exhibit long-range dependence. Our study shows that, at some loads TCP can induce the appearance of self-similarity. In particular, when load is low and loss is rare, traffic looks Poisson. When load is high and the network is overloaded, TCP congestion control can smooth out the burstiness of the aggregate stream so that traffic at the bottleneck tends to Poisson. However, when load is intermediate and the network is prone to occasional bouts of congestion, as is typical of many networks, traffic can become self-similar. Moreover, factors such as round trip time and number of streams passing through the bottleneck can cause the network to become congested at different loads, and consequently affect the range of load over which self-similarity can be observed. The impact of one self-similar TCP stream has also been observed. In particular, our study show that if one or more streams passing through the bottleneck is self-similar and the aggregate flow does not exceed the capacity, traffic observed at the bottleneck will also be self-similar.

I. INTRODUCTION

Recent studies have shown the presence of long-range dependence or even self-similarity in Ethernet LAN traffic [5], World Wide Web traffic [1], Wide Area Network traffic [11], etc. The issue of self-similarity has also been addressed in various studies from many aspects including its impact on network performance [9], modeling techniques [8, 9], and causes of the appearance of self-similarity [1,10,13].

Since the pioneering work on self-similarity of network traffic by Leland et. al., many studies have attempted to determine the cause of this phenomenon. Initial efforts focused on application factors. For example, Crovella and Bestavros [1] investigated the cause of self-similarity by focusing on the variability in the size of the documents transferred and the inter-request time. They proposed that the heavy-tailed distribution of file size and "user think time" might potentially be the cause of self-similarity found in Web traffic.

Alternatively, a few studies have considered the possibility that underlying network protocols such as TCP could cause or exacerbate the phenomenon. In particular, Peha [10] first showed that simple ARQ mechanisms could cause the appearance of self-similarity in congestible networks, but he did not examine the ARQ mechanism in TCP. Veres et al. [13] later showed that TCP could sometimes create self-similarity in an individual TCP stream. Interestingly, in some circumstances, aggregate traffic through bottleneck tends toward Poisson while individual streams remain self-similar, presumably because congestion control mechanisms tend to keep the aggregate throughput close to the capacity whenever load exceeds the capacity. However, the work was based on the assumption that load is infinite (heavy load), which is obviously not sustainable in real networks. In some cases, the heavy-load assumption leads to useful approximations of a congested network, but as this work will show, heavy load produces qualitatively different results from load that is just below network capacity, so observations based on heavy load can be misleading. Finally some researchers [2, 4] considered the effect of TCP in networks with a Bernoulli loss model which randomly drops packets regardless of whether there is congestion in the system or not. Since losses in wired networks are typically due to congestion, and therefore not Bernoulli, this model does not apply to typical wired networks, where self-similarity has been observed [5].

In this work, we use simulations (OPNET 8.0) to show that TCP can induce self-similarity in network traffic even when input from the application layer is strictly Markovian. Our work assumes finite load and uses a full implementation of TCP Reno from the OPNET library. We consider a simple topology where a bottleneck link is subject to bouts of congestion and believe that this potential for congestion is central to the phenomenon. We also show that the timescales over which self-similarity can be observed depend on round-trip time (RTT) and the number of simultaneous TCP sessions.

This paper is organized as follows. Section II describes the network model and network configuration. In section III, we present and discuss our results. Possible future work and concluding remarks are presented in section IV.

II. NETWORK MODEL

We consider a simple network model with N clients and N servers connected by a bottleneck link. Each client is connected to a common switch where all the traffic will be multiplexed and travel through two links in series, with capacities of 100 Mbps and 1.54 Mbps and the MTU of 1500 bytes. The latter link is a potential bottleneck with a finite buffer that can hold up to 10 packets. The total link delay of each client-server pair is set to 5 ms. in all simulations except for the part that investigate the effect of round trip time. Source queue is assumed to be infinite, i.e., infinite buffer.

Fig 1 shows example of a three-client model. Each client generates a packet of size 500 Bytes that is passed to the TCP layer according to a Poisson process with rate λ. There are a total of N streams (TCP sessions) traveling the same direction,

This work is supported in part by the National Science Foundation under Grant NCR 9706491 and by OPNET Technology.

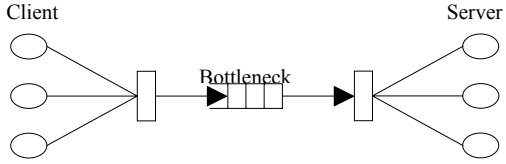

Fig. 1: Three-client network model

III. SIMULATION RESULT

and no traffic in the opposite direction. Each session lasts for 2 hours.

A. Signs of Self-Similarity

We simulate a network with three streams, as shown in Fig. 1, and vary total arrival rate at the application layer from 0.90 Mbps to 1.65 Mbps in order to examine the effect of load on self-similarity of network traffic. Throughout this paper, we measure the degree of self-similarity by plotting a sample variance of average throughput against different averaging duration (m) on a log-log scale (Variance-Time plot or V-T plot). Because variance of self-similar process decays slower than the reciprocal of m, the slope (β) of a V-T curve of a self-similar process is flatter than -1; $-1 < \beta < 0$. Hurst parameter, which is a measure of degree of self-similarity, is related to β as follows: $H = 1+\beta/2$. Although variance analysis technique is not the most accurate method of estimating H, it is sufficient when trying to determine whether the time series of interest is self-similar or not [12].

Fig. 2 shows a sample variance of average throughput from an individual stream as a function of m, where m ranges from 10 milliseconds to 100 seconds. Our first observation is that a network undergoes four qualitatively different kinds of behaviors, which can be seen at four different ranges of load. The strong qualitative differences between these four phases can be seen clearly from the V-T curves and graphs of queue length versus time. We categorize the four different behavoirs as follows.

1. Uncongested phase: Total traffic is sufficiently low that queues rarely form at the sources, see Fig. 3a, and congestion never occurs. When a network is in this phase, TCP will not affect the traffic flows, so TCP traffic resembles application traffic, which is Poisson. Hence, slope of a V-T curve is -1 as shown in Fig. 2 when load is 0.90 Mbps.

2. Sporadically congested phase: At greater loads, the network alternates between congested and uncongested periods. When the network is congested, queues can build up at the sources, as shown in Fig. 3b. Throughput fluctuates wildly during congested period and become rather smooth when there is no congestion, so variance of throughput remains high even when observing over large timescales. As a result, traffic appears self-similar with a slope between 0 and -1 on a V-T plot.

3. Always congested phase: At higer loads, it is possible for a network to remain congested most of the time, so that TCP is almost always limiting the flow of traffic. Throughput at the IP layer is roughly at the maximum and is limited by the TCP congestion control mechanism. Throughput at the application layer equals arrival rate at the application layer, so source queues rarely empty out, as shown in Fig. 3c. Traffic in this phase appears self-similar for a finite range of timescales (*pseudo self-similar* [4,7]) between tens of milliseconds to a few seconds. Slope of a V-T curve is between -1 and 0 for timescales below some threshold and becomes -1 at greater timescales.

4. Overloaded phase: Network is always overloaded when total load is consistently higher than the maximum capacity that the network can handle. Source queues are always increasing, as shown in Fig. 3d, so the arrival process is irrelevant. As a result, the system becomes deterministic because the only source of randomness has been removed. This is the heavy-load case addressed in [13]. It is unlikely to occur over extended periods in real networks, and as our results show, behavior of traffic in this phase is significantly different from the others. In particular, in the scenarios we simulated, each source takes turns increasing and decreasing its transmission rate in a deterministic manner so traffic show periodic patterns. Consequently, the throughput of individual stream fluctuates wildly in some periods and stay smooth in other periods and the burstiness remains even when m is large. Hence, traffic remains self-similar with slope of less than -1 even at a timescale of 100 seconds.

TCP can indeed induce self-similarity on individual TCP stream. However, what is more important is the behavior of aggregate traffic since it is what most of the network components such as switches, or routers see. Fig. 4 shows the V-T plot of aggregate traffic under the same network configuration. Interestingly, as load increases, slope of aggregated traffic approaches -1 and traffic exhibits self-similarity for a shorter range of timescale, i.e., only from tens of milliseconds to hundreds of milliseconds.

More specifically, when network is not overloaded, behavior of aggregate traffic resembles behavior of traffic from a single source. Since an individual TCP stream in the uncongested phase is Poisson, the aggregation of multiple streams is also Poisson. Hence, the V-T curve has a slope of -1. When traffic is in the sporadically congested phase, the superposition of multiple streams yields a traffic pattern that is similar to that of a single stream. That is, the throughput fluctuates wildly when the network is congested and becomes relatively smooth during uncongested period. Hence, aggregate traffic also exhibits self-similarity over all timescales of interest. Finally, an Individual stream appears pseudo self-simliar when network is in the always congested phase, and the aggregation of multiple pseudo self-similar streams gives rise to pseudo self-similar traffic as well.

When the network is overloaded, TCP congestion control mechanism limits the sources' transmission rate. While traffic from an individual stream appears bursty, the aggregated rate is kept relatively steady at 1.54 Mbps or the bottleneck capacity. As a result, the aggregate tends to smooth out very quickly as m increase and consequently the slope of V-T curves tends toward -1. This is consistent with the results described in [13].

B. Effect of Number of Streams

Section III-A shows that TCP can cause the appearance of self-similarity when three streams pass though a bottleneck, and

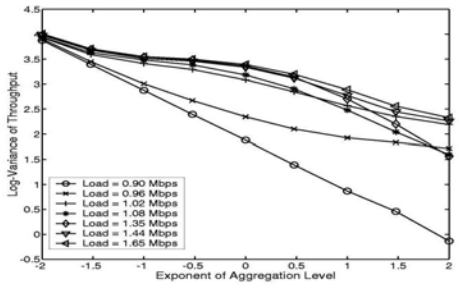

Fig. 2. Effect of load on an individual TCP stream

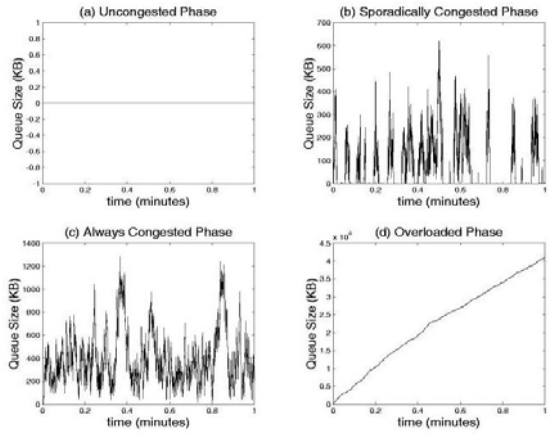

Fig.3. Sum of source queues from four different scenario

that there are four different behaviors at different ranges of load. In this section, we show that the same is true when there are more than three streams. We repeat the simulation of a network in Fig. 1, but with N = 5 and 10. Using the same criteria in section III-A to differentiate the phase of the traffic, our results show that a network with 5 or 10 streams also undergoes the same four distinct phases. However, phase transitions occur at different loads. Table I shows a phase transition table for N = 3, 5 and 10. The four phases are labeled as p1 to p4, corresponding to the network in the uncongested phase to the overloaded phase, respectively

In particular, we observe that with fewer streams, a network starts to experience an extended period of congestion at a lower load in the scenarios we observed. The fact that packets are typically lost in bursts is presumably a factor. More specifically, when there is smaller number of sources and therefore larger data rate per source, it is more likely that there will be enough lost packets in a given stream to trigger full go back N. As a result, there are more retransmissions and congestion is more likely to extend over a longer period than when there are more sources.

As Table I shows, the range of loads in which the network is in the always congested phase is much lower when there are more streams. As a result, in this scenario, when there are many streams, the network is never congested at higher application-layer arrival rates, but it is overloaded at lower rates. Recall that IP-layer arrival rate is roughly the same when the network is in the always congested phase, and a greater application-layer arrival rate can be tolerated without overload by increasing packet size, and thereby decreasing header overhead.

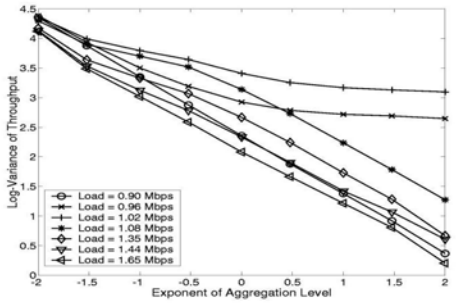

Fig. 4. Effect of load on an aggregated TCP stream.

TABLE I
PHASE TABLE OF NETWORK WITH 3, 5 AND 10 STREAMS

Load	3 streams	5 streams	10 streams
0.96	p1	p1	p1
0.99	p2	p1	p1
1.02	p2	p2	p1
1.04	p2	p2	p1
1.06	p3	p2	p1
1.08	p3	p3	p1
1.10	p3	p3	p2
1.14	p3	p3	p3
1.20	p3	p3	p4
1.35	p3	p4	p4
1.44	p4	p4	p4
1.50	p4	p4	p4
1.55	p4	p4	p4

TABLE II
PHASE TABLE OF NETWORK WITH DELAY OF 5, 15, 25 AND 50 MS.

Load (Mbps)	Total Link Delay			
	5 ms.	15 ms.	25 ms.	50 ms.
0.96	p1	p1	p1	p1
0.99	p2	p1	p1	p1
1.02	p2	p2	p1	p1
1.08	p3	p3	p2	p2
1.20	p3	p3	p3	p2
1.35	p3	p3	p3	p3
1.40	p3	p3	p3	p3
1.42	p4	p3	p3	p3
1.44	p4	p4	p4	p3
1.50	p4	p4	p4	p4
1.55	p4	p4	p4	p4

With a large number of sources, the application-layer arrival rate per stream is smaller, so sources do not always have enough queued data to send very large packets. As a result, overload occurs at a lower load.

V-T plots with five and ten streams scenario are similar to the plots with 3 streams (Fig. 2 and Fig. 4); depending on the phase that a network is in, traffic from individual and aggregate stream show different kinds of behaviors, as explained earlier.

C. Effect of Round Trip Time

We study the effect of RTT by varying the link delay from 5 ms to 50 ms. In this case, congestion occurs at a higher load when RTT is greater. Table II shows a phase table of a network with delay of 5, 15, 25 and 50 ms, where phases are determined using the qualitative criteria described in Section III-A. After a packet is lost due to congestion, the congestion window is halved and is increased once per RTT if the retransmission is successful. When RTT is small, source transmission rate quickly increase to its former level, possibly allowing the congestion to continue. Hence, phase transition of a network where everybody has short RTT tends to occur at a lower load. In other words, traffic exhibits self-similarity at lower load when RTT is small.

In a real network, TCP connections are likely to have widely different RTT. We extend our study by simulating a network

with three streams where each stream has different RTT, i.e., *stream_n; n* = 1,2 and 3 has a delay of 5 ms, 25 ms, and 50 ms, respectively. We refer to this as heterogeneous RTT scenario.

Each stream exhibits the same four behaviors described earlier at four different ranges of load. However, at a given arrival rate, streams with different RTTs may be in different phases. In particular, if *L* is the load that *stream n* moves from one phase to the next, *L* is smaller if the other streams that *stream n* competes with have shorter RTTs. This is consistent with the fact that TCP's congestion control mechanism gives a smaller throughput to a stream that must compete with other streams that have smaller RTTs [3]. For example, Fig. 5 shows a V-T plot of three streams with heterogeneous RTTs when arrival rate ranges from 1.08 Mbps. to 1.50 Mbps. According to Figure 14a, *stream 1*, whose delay is 5ms, experiences no congestion (slope of -1), while the other two streams experience occasional bouts of congestion and appear long-range dependent with slope between 0 and -1. Figure 5b shows that traffic from *stream 3* exhibits pseudo self-similarity while the other two streams are self-similar. At greater loads (Figure 5c-d), all streams appear pseudo self-similar.

The variance of the aggregate traffic, on the other hand, is a function of the variance of all the individual streams and is dominated by the stream with the highest variance. In particular, when total load is less than the network capacity, the slope of an aggregate traffic resembles the one with the largest variance and can possibly exhibits self-similarity if one of the streams is self-similar. Note that if one stream is self-similar, it will inevitably have the greatest variance for larger timescales. However, if load exceeds the capacity, aggregate throughput is limited by TCP congestion control mechanism. Consequently, the slope of a V-T curve of the aggregate stream approaches -1 as load exceeds the capacity. See Fig. 5d.

IV. CONCLUSION

We showed that TCP could indeed cause the appearance of self-similarity in network traffic even with Markovian input at the application layer. In particular, we observed four different kinds of traffic behavior at four different ranges of load in a network where all streams have identical RTT. This is true regardless of the RTT and the number of TCP streams, although the phase transition may occur at different loads with different RTTs and different numbers of streams.

Traffic is Poisson when load is low such that congestion never occurs. At greater loads, the network alternates between congested and uncongested periods, and TCP can induce the appearance of self-similarity over a wide range of timescales. When the network is always congested, TCP traffic appears pseudo self-similar. Finally, when load exceeds the capacity so the network is overloaded, TCP congestion control holds aggregate throughput steady at the network capacity so aggregate traffic exhibits short-range dependence, while each individual stream remains self-similar.

When RTT of each stream is not identical, the level of congestion each stream experiences can be very different, though they go through the same bottleneck. Aggregate traffic of heterogeneous RTT streams tends to smooth out as load exceeds network capacity. However, when the network is not overloaded, aggregate traffic can appear self-similar if one or more streams appear self-similar since variance of the aggregate roughly equal the sum of the individual variances.

Given that TCP can create self-similarity in network traffic in a variety of circumstances, it is likely that self-similarity will probably be observed in all networks that use TCP, even if future applications are Markovian. It may also be possible to develop different algorithms in future generations of TCP that do not create or exacerbate long-range dependence. We believe that this is an open area for future research.

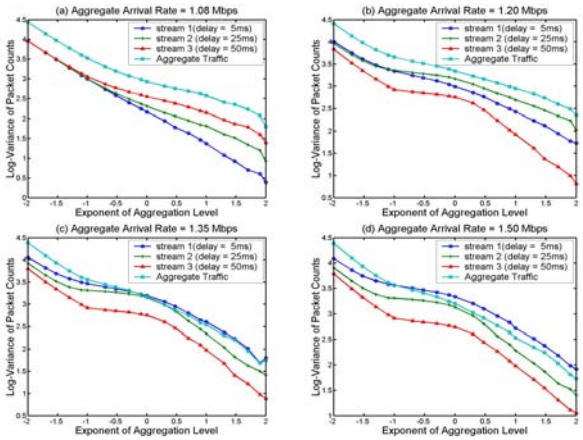

Figure 5. V-T plot of a 3-stream with heterogeneous RTT scenario

REFERENCES

[1] M.E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible cause," *IEEE/ACM Trans. on Networking,* vol.6, pp. 835-846, Dec. 1977.

[2] D. R. Figueiredo, et. al., "On the autocorrelation structure of TCP traffic," *Comp. Networks Journal, Special Issue on Advances in Modeling and Eng. of Long-Range Dependent Traffic*, vol. 40, no. 3, pp. 339-361, Oct. 2002.

[3] S. Floyd and V. Jacobson, "Connection with multiple congested gateways in packet-switched networks, Part1: One-way Traffic," *ACM Comp. Comm. Review,* vol. 21 No.5, pp. 30-47, Aug 1991.

[4] L. Guo, M. Crovella, I. Malta, "How does TCP generate pseudo-self-similarity?" in *Proc. of MASCOTS '01*, Cincinnati, Ohio, Aug. 2001.

[5] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, "On the self-similarity nature of Ethernet traffic," in *Proc. ACM SIGCOMM,* pp. 183-193, Sept. 1993.

[6] T. Le-Ngoc, S.N. Subramanian, "A Pareto-modulated Poisson Process model for Long-Range dependence traffic," *Computer Communication,* vol.23, pp. 123-132, January 2000.

[7] S. Manthorpe, et. al, "The Second-Order Characteristics of TCP," in *Proc. Of Performance '96,* Lausanne, Oct 1996

[8] A. Ost and R.H. Boudewijn, "Modeling and Evaluation of pseudo self-similar traffic with Infinite-State Stochastic Petri Nets," in *Proc. of the workshop on formal method and telecom.,* pp. 120-136, Sept 1999.

[9] K. Park, G. Kim, and M. Crovella, "On the effect of self-similarity on network performance," in *Proc. of the SPIE International Conf. on Performance and Control of Network System,* pp 296-310, Nov 1997.

[10] J. M. Peha, "Protocols can make traffic appear self-similar," in *Proc. of the 1997 IEEE/ACM/SCS Comm. Networks and Distributed System. Modeling and Simulation Conf.,* pp. 47-52, Jan 1997.

[11] V. Paxson and S. Floyd, "Wide-Area Traffic: The failure of Poisson modeling. *IEEE/ACM Trans. on Networking,* vol. 3, no. 3, pp. 226-44, June 1995.

[12] V. Paxon, "Fast Approximation of Self-Similar Network Traffic," *Technical Report LBL36750*, U. of California, Berkeley, Apr 1995

[13] A. Veres and M. Boda, "The chaotic nature of TCP congestion control," in *Proc. IEEE INFOCOM 2000,* Tel Aviv, Israel, pp. 1715-1723, Apr 2000.