The Elusive Nature of Internet Traffic

Woo-Hyun Kim[1], Kyung-Geun Lee[2]
Hyeong-Koo Park[3], Hyo-Jin Lee[4], Ju-Wook Jang[5]

[1,2]Department of Information and Communications Engineering, Sejong University
[3,5]Department of Electronic Engineering, Sogang University
[4]LG Telecom

Abstract: While most previous research on Internet traffic agrees that the traffic of individual connection on Internet exhibits self-similarity, it is divided about whether the aggregated traffic would also be self-similar or Poisson due to multiplexing gain. In this paper, previous claims about the nature of aggregated traffic are reviewed and a new measurement of recent Internet traffic is performed to identify the factors which influence or even distort what we expect from the Internet traffic. What we have found is that Internet traffic depends on so many factors including application mixture, Web browsers, server's policies, saturation and therefore has elusive nature which is hard to be captured into a mathematical model.

1. Introduction

Analysis and modeling of Internet traffic is essential to the efficient design and effective management of access or backbone networks. For example, estimation of traffic at access networks can be used to derive expected queuing delay or response time for each application with given(or scheduled) capacity of the access line. Traffic at access line can be modeled as a queuing system of M/G/1/PS or M/M/1/PS. The arrival is Poisson(M), the service rate depends on the size of accesses and is log-normal(G) or exponential(M), number of servers is one(only one access line), and processor sharing is assumed since packets from different applications are multiplexed to pass through the access line. What concerns us for the design of access networks is the multiplexing gain: whether aggregation of individual accesses smooths out the resulting traffic like aggregation of Poisson processes or not.

While most previous works agree that sizes of individual access show lognormal distribution leading to self-similar or bursty traffic at the packet level, they are divided about the existence and magnitude of the multiplexing gain. Excluding obvious weekly or daily cycles, it is debatable whether aggregation of traffic at the specified busy hour or minute is Poisson or self-similar. Self-similarity is claimed in the packet level analysis[1] and application level analysis[2]. Willinger et al[1] claimed that self-similar traffic can be constructed by multiplexing a large number of ON/OFF sources that have ON and OFF period lengths which are heavy-tailed. Crovella[2] attributes the self-similarity of Web traffic to the heavy-tailed distribution of sizes of the documents stored in Web servers.

On the contrary, a recent work based on the measurements at Havard

University and Lucent Lab. shows that the Web traffic aggregates like Poisson processes[3]. This paper shows a progress of our effort to identify possible causes for this discrepancy among various claims on the nature of Internet traffic. What we have found is that there exists no single model to suit all the traffic patterns appearing at various points at various time. Previous works on Internet traffic disagree not because they interpret the same measurement differently but because they measure different traffic and trying to generalize their findings.

We show this by identifying some factors which greatly influence the measurements: changing application mixture, changing web browsers, emerging new web application and saturation of traffic. For example, comparing the Internet traffic at U.C.Berkeley in the year of 1993 and that at Sogang University, Korea on Sept. 20 2000, one can note that the WWW has risen from a mere existence to the most dominating application. WWW traffic differs greatly from FTP in that short-lived accesses of small files comprise the most of the traffic while FTP usually initiates the long-lived accesses of large files. Since FTP of large files drives the traffic to be more bursty than the WWW access of small files, the traffic pattern will depend on the relative ratio of two applications. There are also some new emerging applications such as VOD, videoconferencing or VoIP and their traffic characteristics may differ from conventional applications. Identifying the nature of each application and the application mixture at the measuring time is essential to correct interpretation of the measurement.

You and Chandra[4] claim that estimation of aggregated traffic at access line should take care of time-stationarity of the traffic and modeling should be based on only the time-stationary traffic. It is observed that Web traffic shows time-stationarity while FTP shows strong non-stationarity. However, this may not apply to new Web applications which allow users to upload their files onto the Web server. In our experiment with a small Web server at Sejong University, the sizes of pages in the Web server used to range from 10k to 60k bytes and the resulting aggregated Web traffic used to be Poisson. But as some users uploaded files as large as 11Mbytes and other users started to download it, the traffic greatly deviated from the Poisson characteristic. This shows an example where even the same application(Web) may show different traffic characteristic depending on the version or policies(such as an administrative limit to the file sizes uploaded).

The rest of this paper consists of the following. Section 2 briefly summarizes the mathematical background needed for understanding of Internet traffic, the tools and the measuring environment. Section 3 discusses our finding and their possible implications. Section 4 concludes this paper.

2. Mathematical background

For 0 <a<1, 0<b<1, the correlation coefficient between the bandwidth at time t and that at time t+k for self-similar traffic is represented as[5]:

$$C_s(t, t+k) = |k|^{-b} \qquad (1)$$

while that of Poisson traffic is represented as:

$$C_p(t, t+k) = a^{|k|} \qquad (2)$$

For example, if k is 4 and a=b=1/2, the correlation coefficient for self-similar traffic is 1/4 while that of Poisson traffic is 1/16. The higher correlation coefficient of self-similar traffic between the traffic separated by k results in burstier traffic than the Poisson traffic. Averaging the traffic with increasing interval rapidly smooths out the Poisson traffic but the self-similar traffic is hard to average out with increasing intervals. More formally, a zero-mean, stationary time series X is H-self-similar if the following is satisfied when $X^{(m)}$ is m-aggregated series of X.

$$\begin{aligned} X &= (X_t \; ; \; t = 1,2,3,\Lambda\ ) \qquad (3) \\ X^{(m)} &= (X_k^{(m)} \; ; \; k = 1,2,3,\Lambda\ ) \\ X_t &= m^{-H} \sum_{i=(t-1)m+1}^{t,m} X_i \quad for\ all\ m \in N \end{aligned}$$

If X_i denotes the traffic produced by user(or host i) and X denotes the sum of X_i, i=1,2,3,…, N, then the variance of X can be obtained as follows[3]:

$$\begin{aligned} \mathrm{Var(X)} &= \sum_{i=1}^{N} \sum_{j=1}^{N} Cov(X_i, X_j) \\ Cov(X_i, X_j) &= Mean(X_i - \overline{X_i})(X_j - \overline{X_j}) \end{aligned} \qquad (4)$$

If users(or hosts) produce the bandwidth independently at a given time instant, the following will hold:

$$\begin{aligned} &\mathrm{Var(X)} = \sum_{i=1}^{N} Cov(X_i, X_j) = \sum_{i=1}^{N} Var(X_i) \qquad (5) \\ &Since \quad Cov(X_i, X_j) = 0 \quad for \quad i \neq j \end{aligned}$$

The user(or host) behavior for Web traffic can be modeled as ON/OFF where ON refers to the transmission period for a Web page and the OFF refers to either the think time up to start of the transmission for next page or just idle time. If we assume the transmission during ON period is fixed with the bandwidth of c and each cycle is formed by k c's followed by N-k zeros, then the $Var(X_i)$ can be represented as follows:

$$\begin{aligned} Var(X_i) &= E(B_i^{\,2}) - E^2(B_i) \\ &= kc^2 / N - (kc / N)^2 \qquad (6) \\ &= c^2 (k / N - (k / N)^2) \\ if\ \ k << N\ then& \\ &= c^2 k / N = c(ck / N) = cE(B_i) \end{aligned}$$

If k << N(OFF time dominates the ON time), variance of each user(host) will be proportional to the average bandwidth of each user(host)[3]. Combining (6) and (5) we would expect the variance of aggregated traffic (Var(X)) would be proportional to the average bandwidth of the aggregated traffic.

3. Characteristics of Internet traffic

Fig. 1 shows the measured round trip time for 24 hours from Sogang University, Korea to Yahoo.com and U. C. Berkeley, respectively. Ping was used to measure the round trip time with resolution of a second. The round trip time during the period from 11:00 AM to 2:30 PM is high and saturated. The interval corresponds to busiest hours for the access router which connects Sogang University to outside Internet. Fig. 2 show the relationship between the average and the variation of the round trip time samples for Yahoo and Berkeley, respectively. Each dot in Fig. 2 represents the average (x-coordinate) and variation (y-coordinate) for a set of 60 consecutive samples taken at each second.

The dots lying near the x axis come from the interval of 11:00 AM to 2:30 PM, which corresponds to the busiest period. During this period, the access line is driven to saturation with heavy traffic. Since RTT stays highest during this period, the variation is very low. If we take samples from the period which is not so busy, for example, for 2:30PM to 6:00PM, Fig. 2 (b) results. Here the access line is released from extremely heavy traffic to reveal its inherent traffic characteristic. First, the correlation between the average and variance of the RTT is as high as 0.916. Second, the variance increases faster than the average RTT. The relationship between the variance is rather close to $y = kx^2$ than $y = kx$(for some k).

Fig. 3 shows the number of connections per each application layer protocol observed on Sep. 20 2000 at the RTI lab. of Sogang University. WWW is clearly the most dominant application, followed by POP3, SMTP, FTP-DATA, FTP, TELNET. Comparing Fig. 3 with Fig. 4 for University of California, Berkeley LBL lab. in the year of 1993, one can observe that the application mixture has changed greatly over time. For example, the WWW has grown from the mere existence used to the most dominating application. The changing mixture of applications would inevitably change the traffic characteristic at the router connecting the access network.

There seems to exist conflicting claims about the aggregated web traffic. Crovella and Bestavros claim self-similarity explaining that it is caused by the heavy-tailed distribution of file sizes found in most web sites[2]. On the contrary, a recent measurement by Morris and Lin shows that Web traffic is aggregated like Poisson processes[3]. We think this discrepancy can be explained by observing that they gathered traffic from different kinds of Web browsers. The Web browsers in [2] are mostly non-commercial and provide a mere graphic interface to the ftp-oriented traffic (with heavy-tailed distribution of file sizes). One can expect that the characteristic of previous FTP traffic would appear from this kind of Web browsers.

The Web traffic in [3] is taken at November 1998 when commercial Web browsers and new Web applications written for general public instead of a few academic professionals blossom. The distribution of sizes of accessed file would be different in that general public would frequently fetch small-sized files instead of an infrequent fetch of a huge file. People tend to cancel the connection which takes more than expected and rather start a new connection. To suit this tendency, most Web sites would contain more files small enough to hold the attention of users. It is shown that the aggregated Web traffic of 24-hours at the T3(45Mbps) line connecting Havard University as well as that of Lucent Labs at the T1(1.5Mbps) line aggregates like Poisson processes[3].

As discussed in Section 2, combining (6) and (5) we would expect the variance of aggregated traffic (Var(X)) would be proportional to the average bandwidth of the aggregated traffic. The variance of aggregated traffic is shown (in the paper) to be linearly proportional to the average, which is a clear indication of Poisson-like behavior. The variance would be proportional to the square of the average if there is high correlation between traffic sources(refer to equation (4) in Section 2). We had expected similar traffic characteristic when we started to measure the traffic at the access routers dominated by Web traffic since today's Web browsers are mostly commercial and basically similar to the ones used in [3]. What we had found is somewhat different from the statistics in [3].

We measured outgoing Web traffic from a Web server at Sejong University. A java program is written to identify each page view and its associated size, ACTIVE ON time and INACTIVE OFF (think time). Fig. 5 shows the diagram for definitions of the various times for a page view. ACTIVE ON is the duration in which actual transfer of data is taking place while INACTIVE OFF is the interval between disconnection of the last TCP connection for the page and the start of the first TCP connection during which users usually read the page or do other jobs. ACTIVE OFF refers to the gap between the two neighboring TCP connections. Sometimes it is not easy to tell from ACTIVE OFF from INACTIVE OFF and so we choose a time threshold of 20 secs below which is considered ACTIVE OFF. 3,482 pages are viewed during Sep. 28(Friday) 2000 to Oct. 2 (Tuesday) 2000.

Fig. 6 shows the relationship between the number of pages viewed and the number of bytes transferred. Each dot in the figure represents a collection of pages viewed for an hour. The x coordinate denotes the number of pages viewed in an hour and the y coordinate denotes the number of bytes transferred. If the aggregated Web traffic has the Poisson characteristic as observed in [3], the dots will be clustered along a line with high correlation as illustrated in Fig. 8 for an ideal Poisson traffic. But our measurement (Fig. 6) shows some discrepancy from this expectation. The correlation coefficient is as low as 0.67. Especially, the two dots (corresponding to measurement of two hours) deviate greatly from the line and thus lower the correlation coefficient greatly.

To identify the cause of this anomaly, we computed the histogram for the sizes of the pages viewed. The maximum size is as large as 11,112,931 bytes. We checked every original page in the Web server and no page is larger than 100 kbytes. The maximum size is more than 100 times larger than the largest original page residing in the Web server. Tracking down the access log files, we found that some users uploaded large document files into the Web server. Excluding the load of extremely large files contributed by users and some abnormal terminations of page views, we obtained the relationship as shown in Fig. 8 which clearly shows the Poisson characteristic. The correlation coefficient is as high as 0.97.

4. Conclusion

We identified some factors which force the real measurements to deviate from what we expect from the Internet traffic as predicted by previous research. In particular, aggregated traffic at access line is considered in detail. We bridged the gap between the two conflicting claims: one that aggregated traffic is self-similar and the other that it is Poisson. Many factors would have influenced the measurements. We identified some factors including application mixture, different Web browsers, emerging Web applications, and illustrated their influences via measurement on real Internet Traffic. Future work will include synthesis of workload which resembles measured traffic under given environment. We will estimate the sensitivity of the traffic to each parameter.

References

[1] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, " Self-similarity through high variability: Statistical Analysis of Ethernet LAN at the source level," IEEE/ACM Trans. on Networking, Vol. 5, pp. 71-86, Feb. 1997

[2] M. E. Crovella, A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and Possible Causes," IEEE/ACM Trans. on Networking, Vol. 5, No. 6 pp. 835-846, Dec. 1997

[3] R. Morris, D. Lin, "Variance of Aggregated Web Traffic," Proc. of Infocom, 2000

[4] C. You, K. Chandra, "Time Series Models for Internet Data Traffic," 24th Conference on Local Computer Networks, 1999

[5] W. Willinger and V. Paxson, "Where Mathematics meets the Internet," American Society for Mathematics, 1998

[6] M. Molina, P. Castelli and G. Foddis, "Web Traffic Modeling Exploiting TCP Connections' Temporal Clustering through HTML-REDUCE," http://www.comsoc.org/ni/private/2000/may/Castelli.html

[7] Domino protocol analyzer, http://www.em-tek.co.kr

[8] Ecoscope, Compuware Corporation, http://www.compuware.com/products

[9] V. Jacobson, C. Leres, S. McCanne, "tcpdump," ftp://ftp.ee.lbl.gov

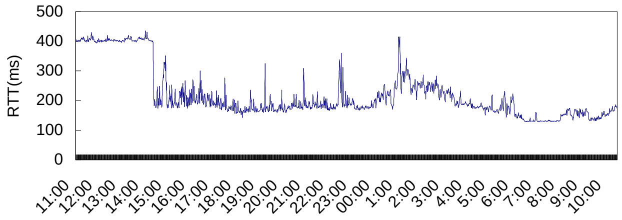
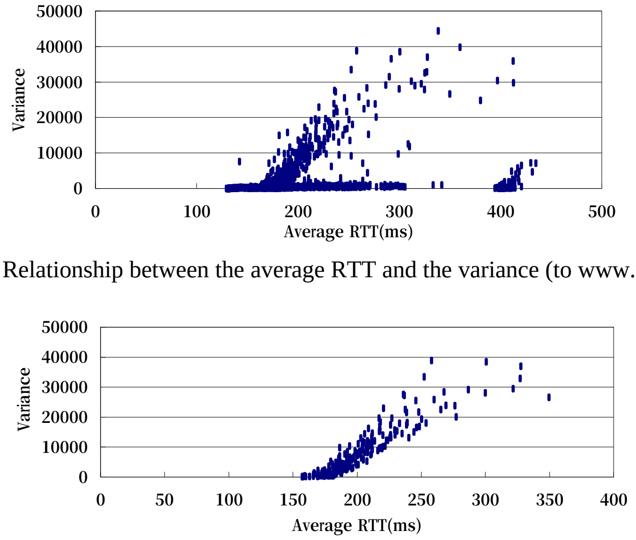


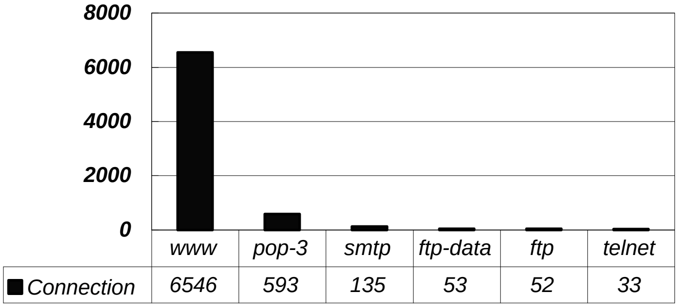
Figure 1 Average RTT from Sogang University (to www.yahoo.com)



(a) Relationship between the average RTT and the variance (to www.yahoo.com)

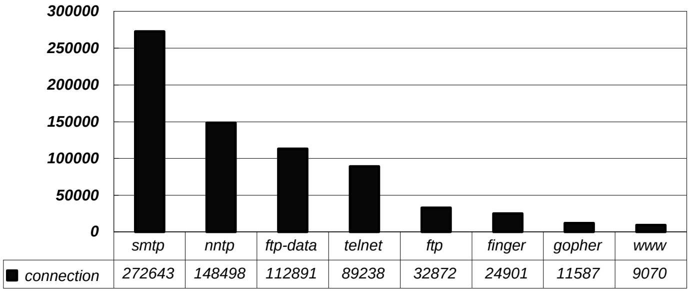
(b) After removal of RTT samples during saturation

Figure 2 Variance vs Average RTT (to www.yahoo.com)



	www	pop-3	smtp	ftp-data	ftp	telnet
■ Connection	6546	593	135	53	52	33

Figure 3 Application mixture in the traffic at RTI lab. of Sogang University (Sep. 2000)



	smtp	nntp	ftp-data	telnet	ftp	finger	gopher	www
■ connection	272643	148498	112891	89238	32872	24901	11587	9070

Figure 4 Application mixture at Berkeley lab(1993)

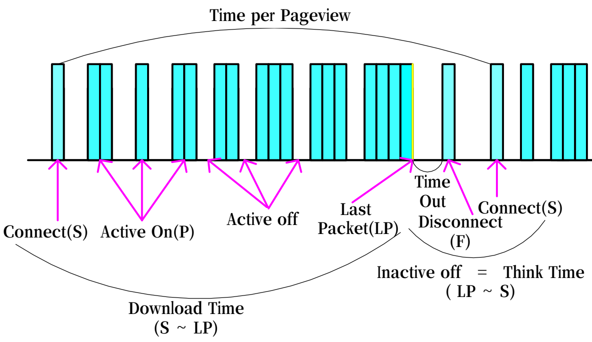


Figure 5 Timing diagram for a page view

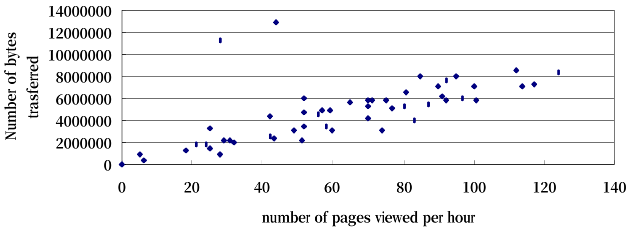


Figure 6 Measured traffic at a web server in Sejong University

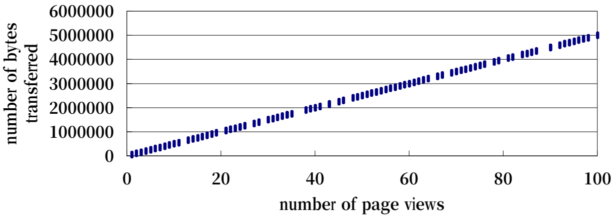


Figure 7 An ideal Poisson traffic

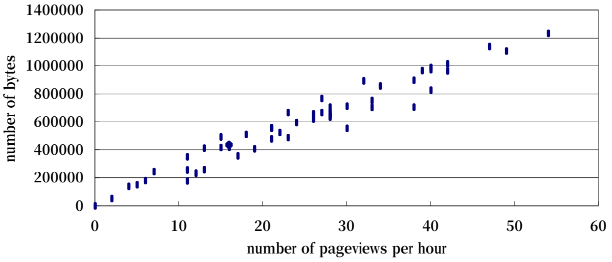


Figure 8 After removal of download for large uploaded files