Variance of Aggregated Web Traffic

Robert Morris
MIT Laboratory for Computer Science
rtm@lcs.mit.edu
Dong Lin
Bell Labs
dong@research.bell-labs.com

This paper appears in the Proceedings of the IEEE INFOCOM 2000 Conference.

Abstract—

If data traffic were Poisson, increases in the amount of traffic aggregated on a network would rapidly decrease the relative size of bursts. The discovery of pervasive long-range dependence demonstrates that real network traffic is burstier than any possible Poisson model. We present evidence that, despite being non-Poisson, aggregating Web traffic causes it to smooth out as rapidly as Poisson traffic. That is, the relationship between changes in mean bandwidth and changes in variance is the same for Web traffic as it is for Poisson traffic.

We derive our evidence from traces of real traffic in two ways. First, by observing how variance changes over the large range of mean bandwidths present in 24 hour traces. Second, by observing the relationship of variance and mean bandwidth for individual users and combinations of users. Our conclusion, that variance changes linearly with mean bandwidth, should be useful (and encouraging) to anyone provisioning a network for a large aggregate load of Web traffic.

I. Introduction

One of the charms of Poisson traffic is its good behavior from the point of view of network engineering. Poisson traffic aggregates well over time, meaning that peaks in the load tend to be canceled out rapidly by succeeding dips. This smoothness around the average means that Poisson traffic can be carried by a network with only modestly more capacity than the average load.

Analysis [1] of Internet traffic has shown that it is not Poisson. Bandwidth used over time turns out to be positively correlated across a wide range of time scales. As a result, peaks and dips do not rapidly cancel each other out, so more capacity is required for good performance than would be needed by equivalent Poisson traffic. In this way Internet traffic is badly behaved.

There is, however, another equally interesting aspect to traffic scaling: aggregation of multiple sources. Poisson traffic aggregates well in this respect: the size of variations in bandwidth increases with the square root of the total bandwidth. This means that as Poisson traffic streams are combined in backbones, the total backbone capacity required increases less than linearly with the total load.

Internet traffic sources fail to aggregate with Poisson-like smoothness in at least one way: the load on backbone links show a strong 24-hour cycle [2]. As a result backbones must be built with enough capacity for the busiest period of the day, rather than just enough for the 24-hour average load. Sources might fail to aggregate well at smaller time scales as well, perhaps due to user habits or coupling at shared bottlenecks. For example, global TCP window synchronization [3] probably causes positive correlation among some connections. Such effects might cause large bandwidth fluctuations within busy periods. Assuming backbone links must have capacities closer to the peak load than to the average, large fluctuations would force a backbone to operate at low average utilization even at peak times of day. Thus the behavior of traffic under aggregation is an important factor in network engineering and economics.

This work investigates the effect of aggregating web traffic on bandwidth variation. The evidence comes from traces of real Internet traffic, analyzed in various ways to focus on different levels of aggregation. The results show that Web traffic aggregates in a Poisson-like well behaved manner.

II. Preliminaries

Our empirical data come from two 24-hour traces of Internet traffic. One trace was taken on a link between Harvard's main campus and its Internet connection. The measured link was a point-to-point 100 megabit Ethernet between two routers. Harvard's Internet link is a 45 megabit T3 line. The measurements were taken over a 24-hour period starting at 3pm EST on April 16, 1998 (a Thursday). The other trace was taken on an Ethernet leading to two of Lucent's T1 Internet connections (since upgraded to T3). This connection serves about 900 Lucent Bell Labs employees. The Lucent trace covered 24 hours starting at 7am EST on Dec 10 1998.

The traces were captured using tcpdump [4] and the Berkeley packet filter [5] on Intel PentiumPro PCs running FreeBSD 2.2. The packet filter reported that it dropped less than 0.01% of incoming packets. Both traces captured each packet's IP and TCP headers, a timestamp, and the packet's total length. Packets in both directions were captured. This work considers only Web traffic, which means that we ignored packets that were not to or from TCP port 80. Roughly half the total traffic on the traced links was Web traffic. The web traffic from both traces is self-similar in the sense that the slopes of the time-variance plots [1] are more than -1: -0.4 and -0.3 for the Harvard and Lucent traces, respectively.

Some parts of this paper appeal to a notion of "user" for purposes of observing the aggregate behavior of different sizes of user populations. We do not know how to attribute packets in our traces to individual human users. Instead we treat local hosts as if they were users. For example, in order to estimate the number

of users who were active during a particular interval of a trace, we count the number of distinct local IP addresses that appeared the source or destination IP header fields in that interval. A local IP address is one on a subnet whose main Internet connection is the traced link. This approach attributes packets received by a local host to that host; this makes sense for traffic arriving at Web clients.

The metric we use to measure bandwidth variation is the variance of bandwidth at 0.1-second intervals. When we wish to compare with bandwidth we use standard deviation (the square root of variance), which has the same units as bandwidth. We consider standard deviation to be interesting because we assume that most network links are engineered to carry the average busy-period load plus a few standard deviations.

We use 0.1 seconds as the interval for variance because Internet routers appear to be able to buffer on the order of 0.1 seconds of incoming traffic. That is, variations in bandwidth at smaller time scales can be smoothed out using buffering. Variations at larger time scales can only be handled using links with capacity greater than the average load (or by discarding data).

We make constant use of the fact that the variance of the sum of independent random variables is equal to the sum of the variances. In particular, if one aggregates a number of uncorrelated sources of bandwidth, the variance of the aggregate will equal the sum of the individual sources' variances. If the sources are statistically similar then the variance will scale roughly linearly with the total bandwidth. Note that the mere fact of aggregating over a shared backbone may cause the sources to become correlated.

For purposes of comparison we define two kinds of abstract traffic. The first, which we call "smooth," has bandwidth whose standard deviation increases with the square root of the average. Poisson traffic is smooth in this sense. The second, which we call "perfectly bursty," has bandwidth whose standard deviation increases linearly with the average. Traffic that doesn't smooth out at all with increasing aggregation is perfectly bursty. For example, Internet traffic is probably perfectly bursty on a time scale of one day since the day/night fluctuation is not likely to smooth out with increasing quantities of traffic.

III. Aggregating Bandwidth

One way to measure how variance scales with aggregated bandwidth would be to measure variance on links supporting differing numbers of comparable users. We can't directly do this since we don't have access to more than a handful of links, and because we don't have a general method for counting or comparing user populations. We can do it for a special case: measurements taken at different times of day on the same link, when different numbers of users are active. These measurements give us a range of bandwidths and corresponding variances.

The upper graph in Figure 1 shows the average bandwidth for each minute of the day from the Harvard trace described in Section II. The lower graph plots the standard deviation during each minute, taken from 0.1-second samples. Figure 2 shows the same data from the Lucent trace. In both cases the standard deviation rises along with the bandwidth from night to day. But how are changes in standard deviation related to changes in bandwidth?

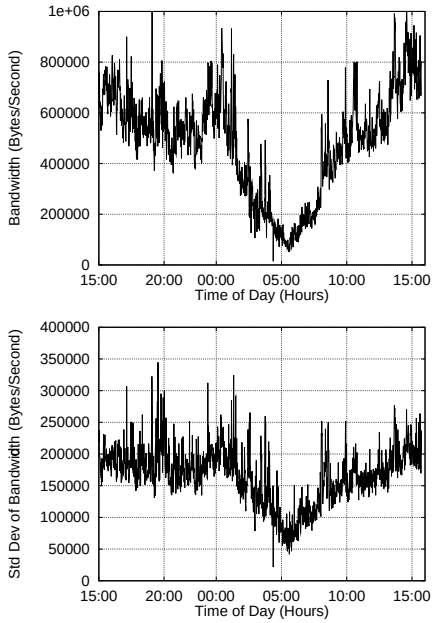

Fig. 1. Per-minute statistics for Web traffic in the day-long Harvard trace. The upper graph shows bandwidth. The lower graph shows standard deviation of bandwidth for 0.1-second intervals.

Figure 3 shows scatter plots with one point for each minute of the day, with x value equal to that minute's average bandwidth, and y value equal to 0.1-second variance during that minute. The solid lines, for comparison, show a linear extrapolation from the variance at the least busy hour, 5am. The average bandwidth of the Harvard trace at 5am is 130000 bytes per second; the variance is 8.1e9. The corresponding values from the Lucent trace are 6352 bytes/second and 277322. The solid lines approximate the relationship of bandwidth and variance for "smooth" traffic (in the sense described in Section II). The dashed curves show the variance of "perfectly bursty" traffic, in which the standard deviation scales with the bandwidth.

The variance in Figure 3 scales almost linearly with the bandwidth. That is, the standard deviation scales with the square root of the total bandwidth. This is the same scaling behavior that Poisson traffic exhibits with respect to aggregation, and suggests that Web traffic is likely to smooth out quickly under increasing aggregation. We explain why this happens in the next sections.

IV. Independence of Users

Two factors might drive the changes in bandwidth and variance in Figure 3: changes in the number of active network users, and changes in the amount of traffic offered by individual users. Figure 4 shows the extent to which the number of users determines the bandwidth. Each point represents one minute in one

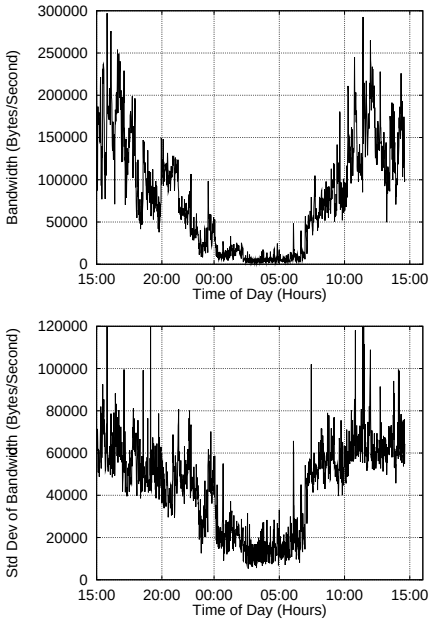

Fig. 2. Per-minute statistics for Web traffic in the day-long Lucent trace. The upper graph shows bandwidth. The lower graph shows standard deviation of bandwidth for 0.1-second intervals.

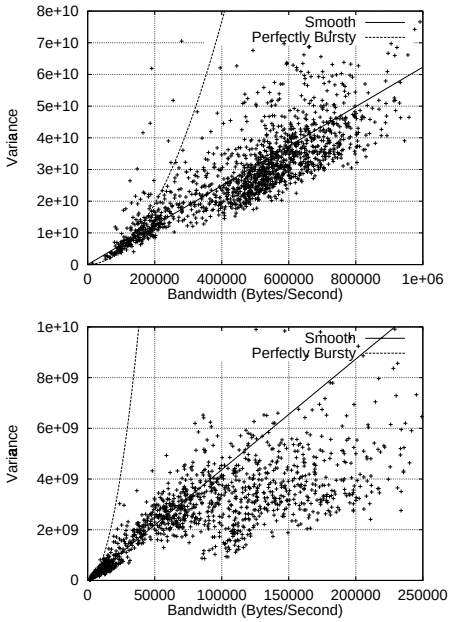

Fig. 3. Scatter plot of bandwidth for each minute versus 0.1-second variance during that minute. The upper graph is from the Harvard trace. The lower graph is from the Lucent trace.

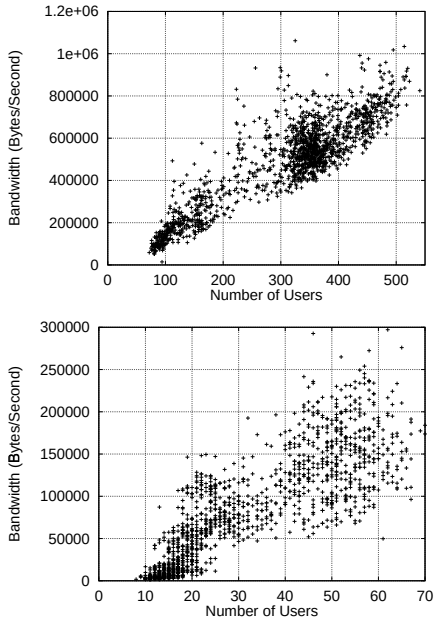

Fig. 4. Scatter plots showing, for each minute of the day-long traces, the number of local hosts active during that minute and the average total bandwidth during that minute. Differences in the number of users account for most but not all of the differences in bandwidth. The upper graph is from the Harvard trace; the lower graph is from the Lucent trace.

of the day-long traces; the x value is the number of "users" (local hosts) active in the minute, and the y value is the total bandwidth during the minute. The correlation coefficient between the two measures is 0.88 for the Harvard trace and 0.84 for the Lucent trace. Thus changes in the number of users explain most of the changes in bandwidth.

A change in the number of users affects the variance in a way that depends on the correlation between users. Changes in the per-user bandwidth also affect per-user variance, which in turn changes the aggregate variance. This section considers correlation, and Section V examines per-user bandwidth and variation.

A. Pair-wise Correlation

Suppose that there are N active network users. Let X_i be a random variable describing the amount of bandwidth produced by the ith user in each 0.1-second interval. Let X describe the total amount of bandwidth from all N users in each 0.1-second interval. The variance of X can be computed from the pairwise covariances:

$$Var(X) = \sum_{i=1}^{N} \sum_{j=1}^{N} Cov(X_i, X_j) \qquad (1)$$

$$Cov(A,B) \equiv Mean((A-\overline{A})(B-\overline{B}))$$

Thus, in order to understand the variance of aggregate traffic, we need to understand the covariance between each pair of users. To make the presentation more intuitive, we will present

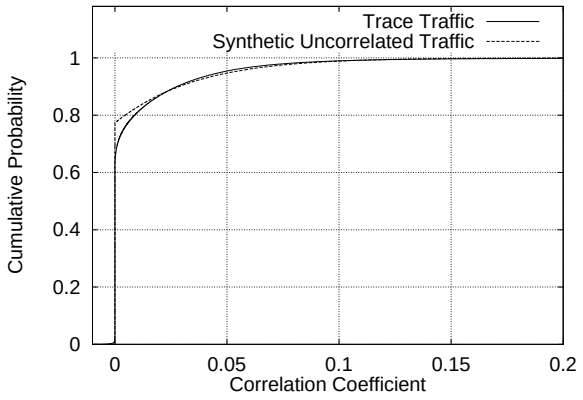

Fig. 5. Cumulative distribution of correlation coefficients between different users' bandwidths. All $N(N-1)$ user pairs are included. The dashed line, for comparison, was synthesized from genuinely uncorrelated sources.

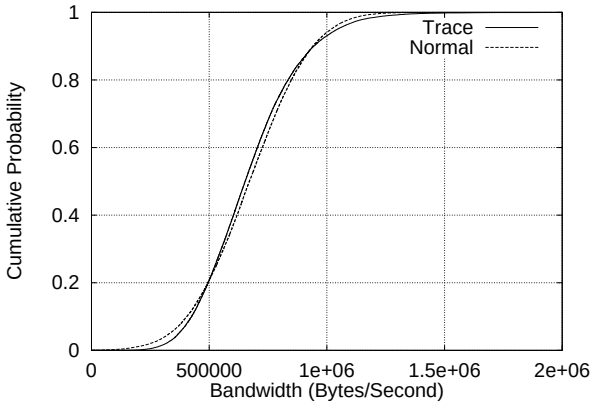

Fig. 6. Distribution of bandwidths from the 3pm hour of the Harvard trace compared with a normal distribution with the same mean and standard deviation. The measurement interval is 0.1 seconds.

pairwise correlation coefficients instead. The correlation coefficient, denoted ρ, is a scaled version of the covariance between two users:

$$\rho = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

We find all pairwise correlation coefficients in the following way. We attribute all web packets to or from a particular local IP address to a corresponding user, as discussed in Section II. To eliminate day/night correlation among users we start with a single hour of the trace, starting at 3pm. We separate the hour into N sub-traces, one for each user. We chop each sub-trace into 0.1-second intervals and calculate the user's bandwidth for each interval. For each pair of users we calculate the correlation coefficient between the users' sub-traces, omitting the correlation coefficient between each user and itself.

Figure 5 shows the cumulative distribution of the resulting $N(N-1)$ correlation coefficients. Though some user pairs have non-zero correlation coefficients, most of this turns out to be due to chance rather than true correlation. To show this, the dashed line shows the correlation coefficient of synthesized uncorrelated sources. The synthesis involved 500 sources, each sending on/off traffic with average "on" time of 4.5 seconds, "off" time of 360 seconds, and inter-packet spacing of 0.06 seconds during the on times. These averages are the same as those measured in the hour trace, and their synthesized distributions were exponential. The good match between the trace and synthetic correlation coefficients suggests that there is very little correlation among users.

Since traffic is generated by a large number of uncorrelated sources, one might expect the bandwidth distribution of the aggregate to be close to normal. Figure 6 show that this is indeed the case for 0.1-second samples. Samples over intervals of 1 and 10 seconds are also nearly normally distributed. This fact allows us to talk about mean, variance, and standard deviation without being misleading. It also suggests that it makes sense to provision network capacity in terms of the mean load plus some number of standard deviations.

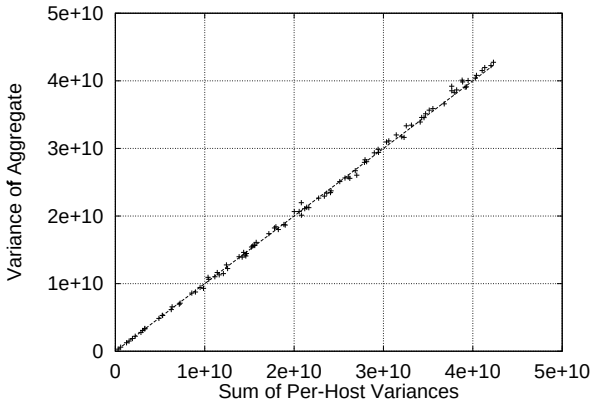

Fig. 7. Comparison of the sum of the per-user variances with the variance of the aggregate for randomly chosen subsets of the 3pm-4pm segment of the Harvard trace. The line $y = x$ is the prediction for uncorrelated users. The subset sizes are uniformly distributed over the total number of users in the trace (2400).

B. Sums of Variances

Section IV-A suggested that the bandwidths from individual users are not significantly correlated. The variance of aggregated traffic should reflect this fact. If users generate bandwidth independently, then $Cov(X_i, X_j) = 0$ if $i \neq j$, so Equation 1 becomes

$$Var(X) = \sum_{i=1}^{N} Cov(X_i, X_i) = \sum_{i=1}^{N} Var(X_i) \qquad (2)$$

That is, the variance of aggregate traffic should equal the sum of the variances of the individual users. One way to test this is to choose various subsets of the users in the trace, calculate the sum of the variances of those users, and compare with the the variance of the subset as a whole.

Figure 7 shows a scatter plot of the results. Each point represents a randomly chosen subset of the users in the 3pm hour of the Harvard trace. Each point's x value is the sum of the variances of those users. Each point's y value is the variance of those users in aggregate. The closeness of the points to the line $y = x$ indicates that Equation 2 holds, and that the users are not noticeably correlated. This lack of correlation helps explain why the variance in Figure 3 changes linearly with the bandwidth, and implies that Web traffic aggregates much like Poisson traffic.

V. Per-User Variance

This section investigates how the variance of per-user bandwidth relates to per-user mean bandwidth. The point is to demonstrate that the variance of an aggregate changes linearly with mean bandwidth even when the mean changes due to individual users consuming more bandwidth. For each user, we calculate 0.1 second bandwidth samples and produce an average and a variance. Figure 8 shows the scatter plots of these per-user variances. Like the variances of the aggregate bandwidth in Figure 3, per-user variance also demonstrates an almost linear relationship to per-user mean bandwidth. We provide a reasonable explanation next.

We model user behavior using ON/OFF cycles. At the beginning of each cycle, the user clicks on a web link and triggers data to be delivered from the remote server. Multiple files might be sent during this ON period. After the web page is loaded, the user reads the document without generating further traffic during an OFF period. We identify an OFF period in the trace when a user is idle for at least 5 seconds. Figure 9 shows the correlations of ON/OFF/transfer size and the user average bandwidth using the Lucent trace. These graphs imply that variation in user bandwidth is mainly caused by different OFF periods and that the average ON periods and transfer sizes tend to have less effect on the bandwidth for most users. Furthermore, 80% of OFF periods are at least 10 times bigger than ON periods. Similar results are obtained from the Harvard trace.

We can further which of transfer size, on period, or off period has the most effect on average bandwidth using a sensitivity analysis. Let B be a user's average bandwidth of a cycle, X be the transfer size, N be the number of bandwidth samples taken in a cycle, and T, T^{on}, T^{off} be the total time, ON period, and OFF period respectively. We evaluate B as a function of X, T^{on}, and T^{off} using samples from the traces. Theoretically, if we fix two of the three variables and vary the other, B should change according to its "sensitivity" to the unfixed variable. The most sensitive variable should produce the biggest variance of B. Such a sensitivity test can be performed using our trace samples presented in Figure 9. To compute the sensitivity for T^{off}, we take a subset of the samples which have similar X and T^{on} values and compute the variance of B. The process is repeated for all possible values of X and T^{on}. The average of these variances is taken as the sensitivity for T^{off}. The same procedure is performed for B and T^{on}. The following table lists the computed sensitivities. For both traces, OFF period has the largest sensitivity. T^{off}'s sensitivity in the Harvard trace is ten times larger than the other two. Lucent samples tend to have large variances due to the small number of samples. There are about 7000 users in the Harvard trace but only 400 in the trace from Lucent. As a result, the number of samples per subset is small and large variances are introduced. These findings helped us derive an argument for Figure 8.

	T^{off}	T^{on}	X
Harvard	3940553	135386	430399
Lucent	5256257	2981678	1253812

We now estimate the variance of one cycle in terms of the bandwidth. Assume ON periods and transfer sizes are fixed and that OFF periods are much larger than ON periods. For simplicity, we further assume that all bandwidth samples during ON periods have the same value for all users, say c. Since transfer sizes and ON periods are fixed, samples from each cycle are just k c's followed by $N - k$ zeros, where k is a constant. Therefore, the variance is

$$\begin{aligned} V &= E(B^2) - E^2(B) \\ &= kc^2/N - (kc/N)^2 \\ &= c^2(k/N - (k/N)^2) \end{aligned}$$

Because $T^{\text{on}} \ll T$ and $k/N = T^{\text{on}}/T$, the variance can be simplified to

$$V \approx c^2 k/N = c(ck/N) = cB$$

The above equation shows that the variance of a user's bandwidth is mostly proportional to the user's average bandwidth. This partially contributes to the similar proportional pattern in the aggregate bandwidth as demonstrated in Figure 3. There are cases in which the average transfer size is monotonic with the average bandwidth, as shown in the upper right graph of Figure 9. This will make the variance to be quadratic with the average bandwidth. But Figures 8 and 9 show that these cases are rare and do not affect the overall shape of the curves.

VI. Related Work

Our techniques and framework owe a lot to investigations of long-range dependence (LRD) in data traffic [1], [6], [7]. LRD studies observe that traffic streams tend to exhibit correlation in time, causing unexpectedly high variation across all time scales. We, in contrast, observe that no similar correlation exists among sources of bandwidth, so that aggregation of sources causes the traffic to smooth out.

Our use of source-level models to explain aggregate behavior is based most directly on recent work explaining long-range dependence [8], [9], [10]. Again, our contribution is to observe burstiness while varying the number of sources aggregated.

We consider only cases in which the traffic stream encounters no common bottleneck. A number of existing studies investigate situations dominated by bottlenecks, queuing and congestion control [11], [12], [13], [14], [3]; the resulting traffic behavior is quite different from that presented here. Our work applies to network links intended to be fast enough that they are not bottlenecks.

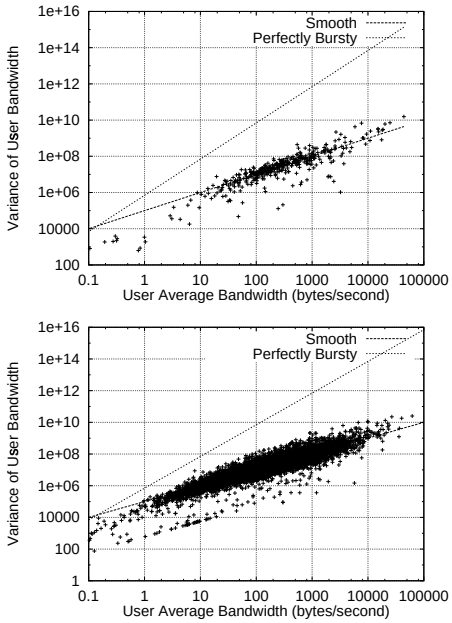

Fig. 8. Per-user variance over average bandwidth from 24 hour traces in log scale. Variances and average bandwidths are calculated from 0.1 second samples for each user. The upper graph is produced from a Lucent trace. The lower graph is from Harvard.

VII. Conclusions

Using measurements from traces of real Web traffic, we present evidence that bandwidth variance changes linearly with the mean. We present two explanations. First, there appears to be little correlation among users in their consumption of bandwidth. This leads to linear variance changes to the extent that changes in mean bandwidth are due to changes in the number of active users. Second, individual Web users consume differing amounts of bandwidth mostly by pausing longer between transfers. This tends to produce variance linear in each user's mean bandwidth. Thus, to the extent that total bandwidth changes are due to users changing the amount of bandwidth they use, the total variance is also linear in the mean total bandwidth.

These results apply only to interactive Web traffic; they are only useful on links engineered to avoid persistent queuing; and they are only useful within the context of a busy hour. Within these limits the results mean that Web traffic gets smoother with aggregation with the same rapidity as Poisson traffic.

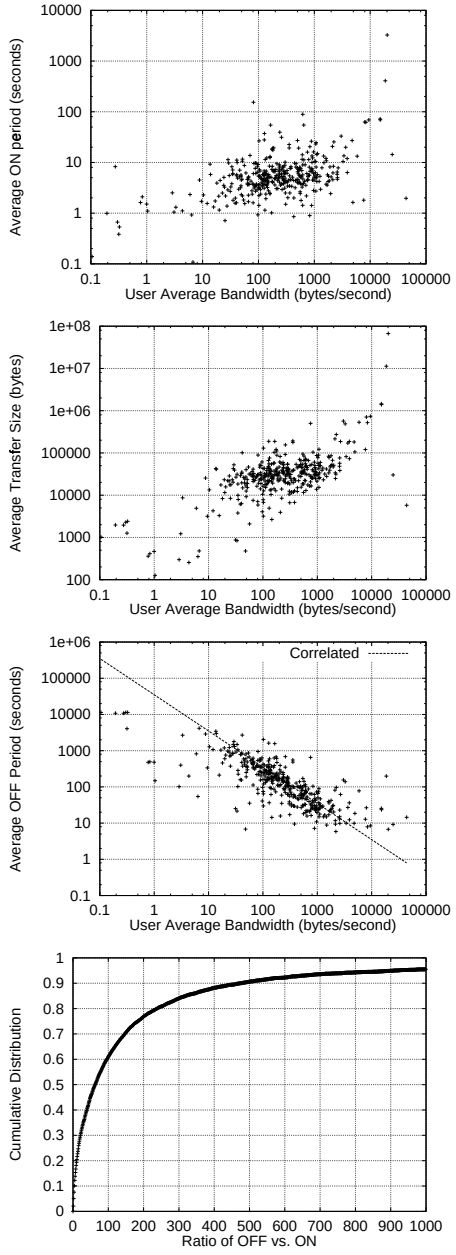

Fig. 9. Correlation of ON/OFF/transfer size over per-user bandwidth. ON periods and transfer sizes do not vary substantially with the bandwidth. There is an inverse linear correlation between the OFF period and the bandwidth. The lower graph shows the cumulative distribution of the ratio between OFF and ON periods. 80% of OFF periods are at least 10 times bigger than ON periods. Graphs are made from the Lucent trace. Similar results are obtained from the Harvard trace.

References

[1] Will Leland, Murad Taqqu, Walter Willinger, and Daniel Wilson, "On the self-similar nature of Ethernet traffic," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, February 1994.

[2] Kevin Thompson, Gregory Miller, and Rick Wilder, "Wide-area Internet traffic patterns and characteristics," *IEEE Network*, November/December 1997.

[3] Scott Shenker, Lixia Zhang, and David Clark, "Some observations on the dynamics of a congestion control algorithm," in *Proceedings of ACM SIGCOMM*, 1990.

[4] Van Jacobson, Craig Leres, and Steve McCanne, "tcpdump," ftp://ftp.ee.lbl.gov.

[5] Steve McCanne and Van Jacobson, "The BSD packet filter: A new architecture for user-level packet capture," in *Proceedings of the Winter USENIX Conference*, 1993.

[6] Henry Fowler and Will Leland, "Local area network traffic characteristics, with implications for broadband network congestion management," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 1139–1145, September 1991.

[7] Ashok Erramilli, Onuttom Narayan, and Walter Willinger, "Experimental queueing analysis with long-range dependent packet traffic," *IEEE/ACM Transactions on Networking*, vol. 4, no. 2, pp. 209–223, April 1996.

[8] Vern Paxson and Sally Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, June 1995.

[9] M. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," in *Proceedings of SIGMETRICS'96*, 1996.

[10] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet lan traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, February 1997.

[11] John Nagle, "On packet switches with infinite storage," RFC970, IETF, 1985, `ftp://ftp.ietf.org/rfc/rfc0970.txt`.

[12] Matthew Mathis, Jeffrey Semke, Jamshid Mahdavi, and Teunis Ott, "The macroscopic behvior of the TCP congestion avoidance algorithm," *ACM Computer Communication Review*, vol. 27, no. 3, July 1997.

[13] Curtis Villamizar and Cheng Song, "High performance TCP in ANSNet," *Computer Communications Review*, vol. 24, no. 5, October 1994.

[14] Sally Floyd and Van Jacobson, "On traffic phase effects in packet-switched gateways," *Internetworking: Research and Experience*, vol. 3, no. 3, September 1992.