Statistical Characterization of Wide-Area IP Traffic

Matthew T. Lucas † *Dallas E. Wrege* ‡ *Bert J. Dempsey* ⋆ *Alfred C. Weaver* †

† Department of Computer Science
University of Virginia
Charlottesville, VA 22903
{matt, weaver}@Virginia.edu

⋆ School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
bert@ils.unc.edu

‡ IBM Corporation
Network Studies Group
Research Triangle Park, NC 27709
wrege@raleigh.ibm.com

Abstract

Background traffic models are fundamental to packet-level network simulation since the background traffic impacts packet drop rates, queuing delays, end-to-end delay variation, and also determines available network bandwidth. In this paper, we present a statistical characterization of wide-area IP traffic based on 90-minute traces taken from a week-long trace of packets exchanged between a large campus network, a state-wide educational network, and a large Internet service provider. The results of this analysis can be used to provide a basis for modeling background load in simulations of wide-area packet-switched networks such as the Internet, contribute to understanding the fractal behavior of wide-area network utilization, and provide a benchmark to evaluate the accuracy of existing traffic models. The key findings of our study include the following: (1) both the aggregate packet stream and its component substreams exhibit significant long-range dependencies in agreement with other recent traffic studies, (2) the empirical probability distributions of packet arrivals are log-normally distributed, (3) packet sizes exhibit only short-term correlations, and (4) the packet size distribution and correlation structure are independent from both network utilization and time of day.

Key Words: Traffic Modeling, Internet Traffic Characterization, Network Simulation, Self-Similar Traffic, Internet Flows.

1 Introduction

Simulation modeling of computer networks is a powerful technique for evaluating the design and performance of network, transport, and application-level protocols. Background traffic models are a fundamental component of packet-level network simulators since the network load drives the packet drop rate, queuing delay, end-to-end delay variation, and available network throughput [7]. Developing background traffic models suitable for use in a large-scale, packet switched network simulation (e.g., an Internet backbone network simulator) is a difficult problem for the following reasons: (1) backbone networks are often inaccessible for measurement and study, (2) the nature of Internet applications, user populations, and user demand is constantly changing, and (3) network traffic is shaped by network switches as well as end-system congestion control protocols.

The goal of the traffic characterization presented in this paper is to support the development of an efficient and accurate background traffic model for WAN simulations. In particular, we are interested in developing packet-level simulations of large-scale IP backbone networks that provide service for large university, enterprise, or small ISP networks. Recent measurement-based studies (see[6, 11, 13] and references therein) have established the *self-similar* nature of network traffic in several contexts and developed techniques to model such traffic. This paper presents the statistical characteristics of network traffic exchanged between the campus network at the University of Virginia, a state-wide educational network, and a large Internet service provider. Our analysis confirms the self-similarity of the campus-level IP packet stream, determines parameter values under different network loads, and analyzes an address-based partitioning of the aggregate stream that is useful for WAN modeling. The results of this analysis can be used to provide a basis for modeling background load in simulations of wide-area packet-switched networks such as the Internet, contribute to understanding the fractal behavior of wide-area network utilization, and give a benchmark to evaluate the accuracy of existing traffic models.

The remainder of this paper is organized as follows: Section 2 describes how the packet traces were collected and gives an overview of the traces used throughout the paper. Section 3 presents the statistical properties of the aggregate packet stream generated by the UVA campus network. We demonstrate that the arrival density function follows a log-normal distribution, exhibits significant long-range dependencies (LRD) over the entire range of network utilization, and the LRD is consistent with LAN traffic. In the context of modeling the campus traffic stream, the findings show that established techniques for generating synthetic streams of self-similar traffic [3, 4, 13] are well-suited for WAN background traffic models. Section 3 also presents the density function and correlation structure of the packet sizes, and finds that the packet sizes have only short-range dependencies with a density that is independent of the network load.

Section 4 presents the arrival density and correlation structure of the substreams obtained by partitioning the aggregate UVA streams along destination IP addresses. Characterizing substreams is important since, in the context of simulation, an accurate partitioning of background traffic introduced into the WAN backbone is crucial. In our partitioning, a few substreams comprise the majority of the aggregate traffic, and these substreams exhibit statistical properties similar to those of the aggregate stream. However, the partitioning also creates very light substreams, e.g., ones contribute less that 3% of the aggregate packet stream, and these substreams do not exhibit the same degree of self-similarity as the larger substreams and aggregate streams. These findings are not surprising in that the amount of source aggregation is an important factor in determining the presence and degree of self-similarity for a traffic stream [11]. For WAN modeling using a self-similar traffic model, the findings show that very light substreams are difficult to characterize using the properties of the aggregate stream and may require separate treatment. Conclusions and future work are discussed in Section 5.

2 Collection of Wide-Area IP Packet Traces

The analysis presented in this paper is based on 90-minute samples from a week-long trace of nearly one billion IP packets exchanged between the University of Virginia's campus network (UVAnet), the Virginia Educational and Research Network (VERnet), and BBNplanet (at the time, UVA's Internet service provider). The network monitor used to collect the trace consists of a powerful workstation[1], a kernel customized to have large network buffers, and a kernel-level packet filter [1]. The network monitor provides a timestamp resolution within 100μsec and an observed drop rate of 0.005% over the entire trace.

Figure 1 shows the experimental setup which consists of three routers and a network monitor interconnected by an Ethernet hub. The VERnet and BBNplanet routers are each connected to three T1 links, while the UVAnet router is connected to UVA's backbone FDDI concentrator. The filter is configured to listen promiscuously on the Ethernet and capture all IP packets sent between the UVAnet, VERNet and BBNPlanet routers. The filter captures the IP header and saves the IP source, IP destination, timestamp, and size of each packet to disk. After compression, approximately six bytes are saved per packet. The week-long packet trace (consisting of 6GB of data) is publicly available at [2].

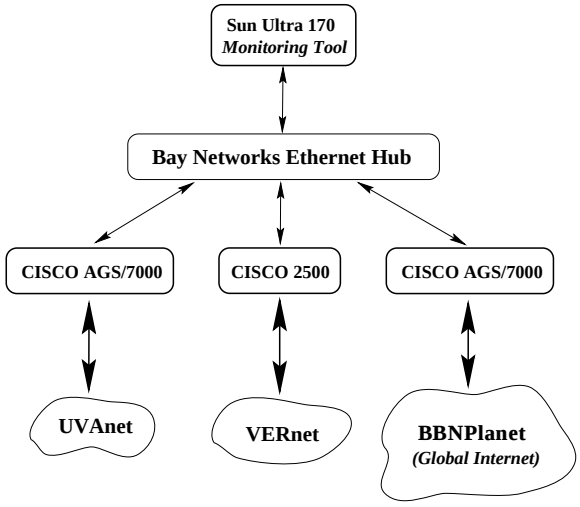


Figure 1. Experiment Setup

Figure 2 depicts the nine-day packet trace captured by the packet filter. The figure plots the number of packets exchanged between the three networks per 100-second interval as a function of time. There are two periods where the monitor workstation went off-line. The first period occurred between 8PM Wednesday and 8AM Thursday due to a disk problem, and the second failure occurred at 11PM on the second Tuesday due to a campus-wide power outage. Two interesting observations about the data are: (1) the ratio of the peak to the minimum data rate is approximately 8:1, which is bursty at this timescale, and (2) the packet rate is cyclical with periods of low utilization occurring around 5AM and peak utilization occurring around 4PM.

3 Data Analysis

This section presents the packet size distribution, arrival correlation, and arrival density of the aggregate traffic generated by UVAnet (i.e., the traffic leaving UVAnet destined for either BBNplanet or VERnet). The analysis focuses on the 27-hour period highlighted in Figure 2. Figure 3 shows the

[1] Sun UltraSparc Model 170, 100MB RAM, 8GB HD running Solaris 2.5

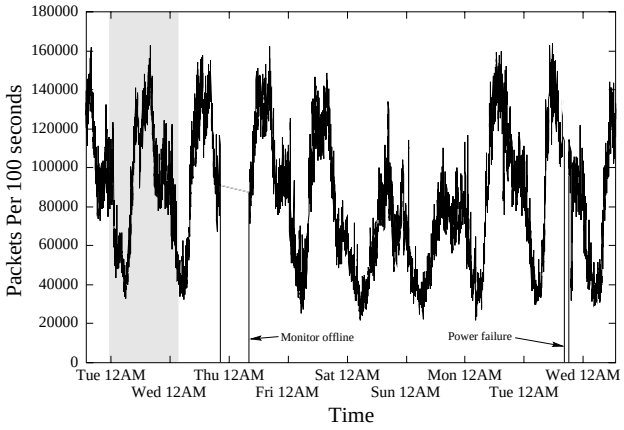


Figure 2. Packets per 100 seconds for 9 day packet trace.

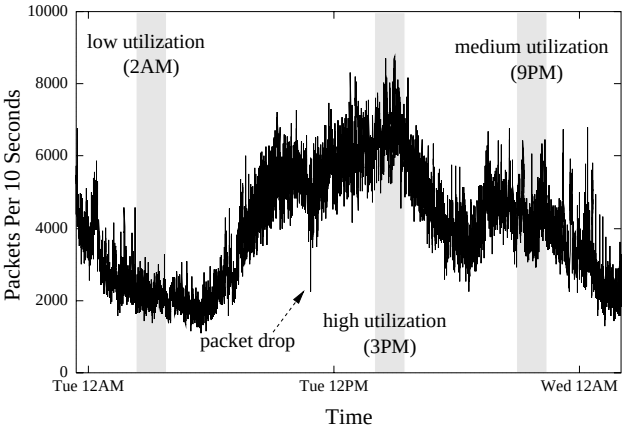


Figure 3. Packets per 10-second interval for 27 hour packet trace.

packets generated by UVAnet per 10-second interval during this period. As indicated in Figure 3, only a single network monitor fault occurred during the trace; just before 12PM, the monitor timed out for ten seconds and dropped $9,784$ packets.

The statistical analysis focuses on the three 90-minute intervals highlighted in Figure 3. These intervals, namely the 2:15AM – 3:45AM ("2AM trace"), 2:00PM – 3:30PM ("3PM trace"), and 9:00PM – 10:30PM ("9PM trace"), were selected because they correspond with periods of low, high and medium network utilizations, respectively, and because the arrival processes are stationary over the duration[2]. Although only three traces from a single 27 hour trace are presented, the analysis here is consistent with and representative of that done with other data sets from the week-long trace.

3.1 Packet Sizes

We first present the density and correlation structure of the packet sizes for each trace. Figure 4 shows the empirical probability distribution of packet sizes for the 2AM, 3PM and 9PM traces. The density is presented on a logarithmic scale to highlight that a small number of packet sizes dominate the trace. In particular, approximately 75% of the packets are either $40 - 44$, or 552 bytes in length. Inspection of

[2] Note that to evaluate the correlation structure of the packet arrivals, the process must be stationary.

the distribution also reveals "spikes" at $55, 60, 75, 144, 576$ and 1500 byte packets, accounting for 12% of the packets. A key observation in Figure 4 is that the densities are nearly identical for all three traces, which shows the distribution of packet sizes is independent of network utilization.

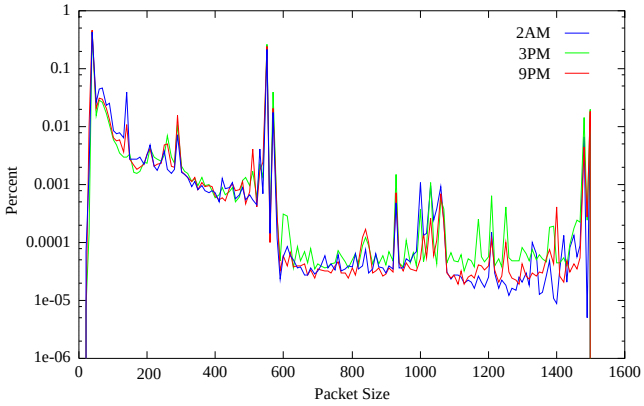


Figure 4. Probability density function of packet sizes.

We next consider the correlation structure of the packet sizes. For a random process $\{X_i\}_{i=0,1,\ldots,N}$ with sample mean $\overline{X}$ and sample variance of S^2, the autocorrelation function r can be estimated for all lag k as follows:

$$r(k) = \frac{\sum_{i=1}^{N-k}(X_i - \overline{X})(X_{i+k} - \overline{X})}{(N-k)S^2} \qquad (1)$$

Figure 5 gives the autocorrelation $r(k)$ of the packet sizes plotted as a function of the lag k for each trace. Since the tail converges rapidly to 0, we can conclude that packet sizes are not correlated; i.e., the size of packet x_i has a negligible influence on the size of packet $x_{i+1}, \ldots, x_n$. The lack of correlation can be explained by the nature of statistical multiplexing in IP networks. That is, packet sizes are most often highly correlated as they are generated by the application. However, as the network statistically multiplexes a large number of independent connections, the correlation diminishes. For example, Figure 5 shows that the $9PM$ trace has the most correlation, and the $3PM$ trace has the least correlation. The correlation analysis shows that packet sizes can be faithfully modeled for a large campus network by independently choosing a packet size using the empirical density function shown in Figure 4.

3.2 Packet Arrival Correlation Structure

Next we consider how packet arrivals at the network monitor are correlated over time. Accurately modeling the arrival correlation of the traffic stream injected by the campus network into the WAN backbone is critical since packet bursts dramatically affect the packet drop rate, variation in network transmission delay, and available network throughput within a wide-area network simulation.

The time-dependent properties of the UVAnet streams are shown in Figure 6 by plotting the autocorrelation function of packet arrivals per $1\ ms$. In contrast to the packet size correlation, Figure 6 shows the correlation structure of packet arrivals is hyperbolically decaying, suggesting that the streams have long-range dependencies (for a review of this subject, see [5]). The important property of an arrival process with long-range dependencies is that the arrival burstiness is similar, independent of the time scale in which it is viewed (so-

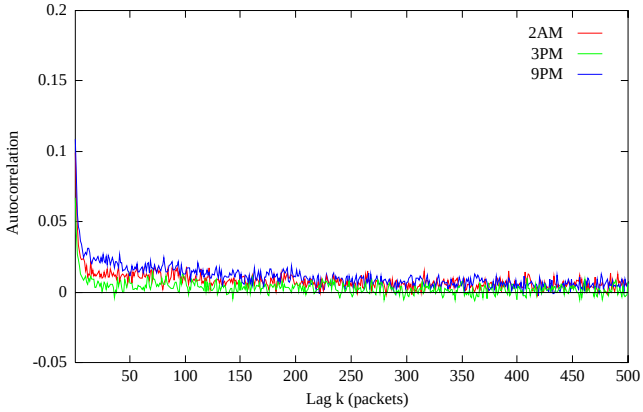


Figure 5. Autocorrelation function for packet sizes.

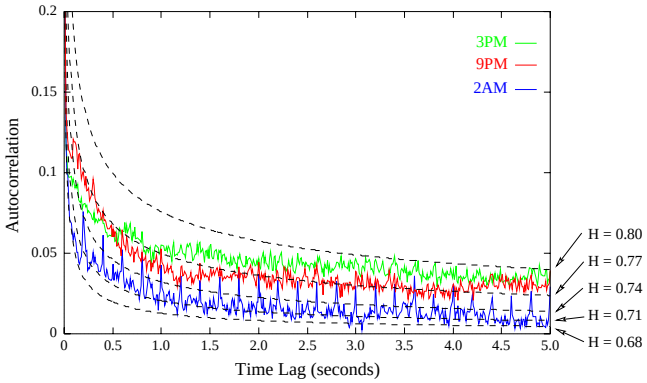


Figure 6. Autocorrelation function for the 2AM, 3PM and 9PM traces.

called *self-similar* processes). Formally, a stationary process $\{X_i \mid i = 0, 1, \ldots, \infty\}$ and its associated *aggregated arrival processes* $\{X_i^{(m)} \mid m = 1, 2, \ldots, \infty\}$ given by:

$$X_i^{(m)} = 1/m \sum_{k=im}^{i(m+1)-1} X_k \qquad (2)$$

is *exactly second-order self-similar* if the autocorrelation $r^{(m)}(k)$ of each aggregated process is given by [11]:

$$r^{(m)}(k) = r(k), k \geq 0 \qquad (3)$$

and the variance is given by [11]:

$$\text{Var}(X^{(m)}) = Var(X)m^{-2(1-H)} \qquad (4)$$

The degree of self-similarity is expressed by the Hurst parameter H in equation (4). H varies between 0.5 and 1, where a larger value indicates a higher degree of self-similarity. For a short-range dependent process, such as the Poisson-based models in [9, 14], the Hurst parameter will be approximately 0.5; thus, by (4), the correlation of a Poisson process will fall off as $1/m$ where m is called the *aggregation level*. Using the reference curves in Figure 6 we see that the correlation structure of the traces correspond to self-similar processes with H between 0.70 and 0.80; thus, they can not be accurately modeled with a Poisson-based process. These results are consistent with studies showing the self-similarity

of LAN traffic which have estimated the Hurst parameter as high as 0.82 [11].

In order to better evaluate the self-similar nature of the traffic, we consider log-variance plots for the three traces in Figure 7. Log-variance plots show the degree of burstiness of an arrival process over multiple time scales by plotting the $\log_{10}$ of the normalized variance of the aggregated arrival process $X^{(m)}$ against the $\log_{10}$ of the aggregation level, m. In contrast to a short-range dependent or Poisson process (i.e., where $Var(X^{(m)})$ falls off as $1/m$), Figure 7 shows that the variance of the arrivals for all three traces decay slowly, in proportion to a self-similar process with $H = 0.65$ for small aggregation levels, and asymptotically as a self-similar process with $H = 0.8$.

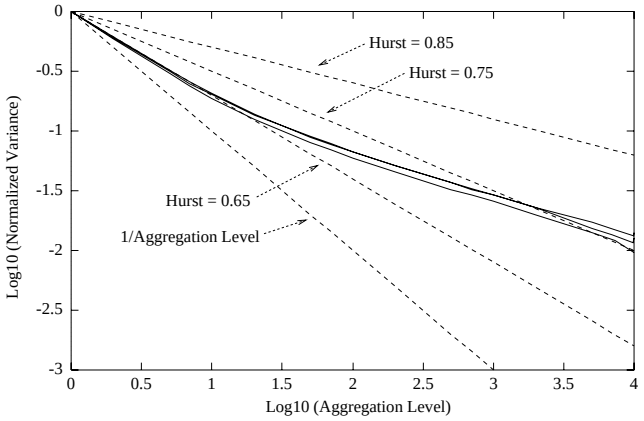


Figure 7. Log-variance of empirical aggregate traces.

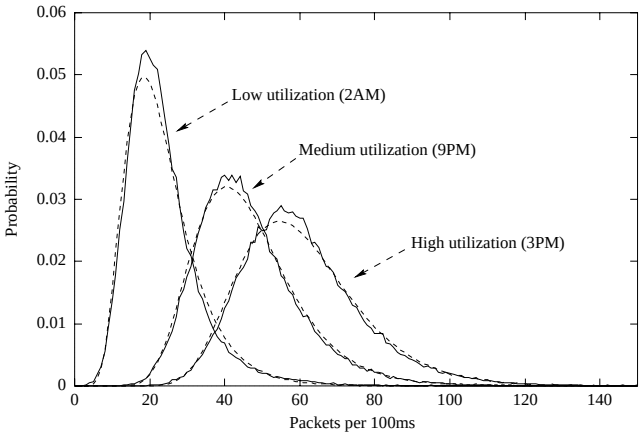
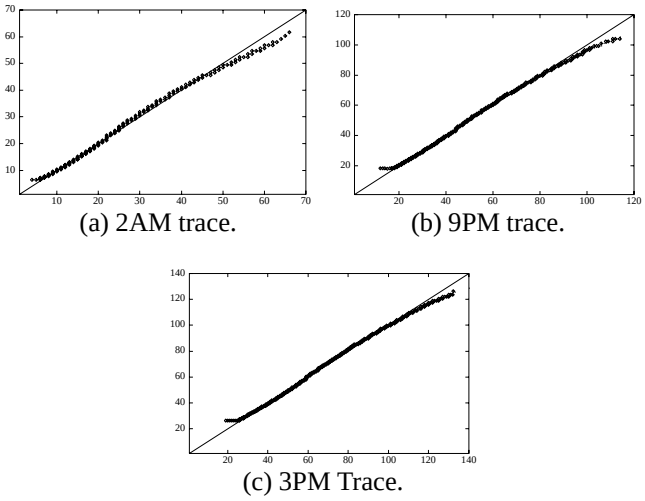


Figure 8. Histogram and log-normal fit of packet rates at low, medium, and high network utilizations. Empirical traces are shown as solid lines, while their log-normal approximations are depicted as dashed lines.

An important class of processes that can model fractal traffic are so-called self-similar models such as fractional Gaussian noise [12] and fractional ARIMA processes [3]. Models that approximate fractional Gaussian noise [8, 10, 13] are attractive for their computational efficiency and simplicity (most of these models only require the Hurst parameter as input). Typically, these traffic models generate sample paths that are normally distributed. Thus, the sample path must be converted to match the density of the empirical traces. Figure 8 shows the distribution of packet arrivals for UVAnet per $100\ ms$ interval. The solid lines give the empirical probability density of the traces, while the dashed lines represent analytical log-normal distributions that are fit to the empirical densities using a maximum likelihood estimator (MLE). As the figure shows, the fit appears to be good except for the deviation at the distribution peaks. To better evaluate the goodness of the fit, we show Q-Q plots in Figure 9 that plot the quantiles of the empirical data against the quantiles of the fitted distribution. Figure 9 shows that the log-normal distribution is very good across the entire range of the distribution except for the tail at the right where the log-normal approximation overestimates the empirical data. This suggests that existing self-similar models may be useful in matching our empirical data since normally distributed processes can be easily transformed to a log-normal form.


(a) 2AM trace. (b) 9PM trace.

(c) 3PM Trace.

Figure 9. Q-Q plots of 2AM, 9PM and 3PM empirical traces versus fitted log-normal distributions.

4 Partitioning the Aggregate Packet Stream

In the previous section we considered the traffic of an aggregate stream departing a campus network. This section presents the statistical properties of substreams obtained by partitioning the aggregate traffic along destination network addresses. In the context of our modeling approach discussed in the introduction, characterizing component substreams is important because background traffic models must not only construct an aggregate packet arrival process but must associate a destination campus address or network access point with each packet. Like the aggregate stream, the correlation and density of each substream must be accurately modeled, otherwise network drop and delays within the network simulation will not reflect the performance characteristics of a production network.

We divide the aggregate stream into 14 substreams based on their destination IP addresses. Table 1 gives the network mask used to define the component substreams and the percentage of packets each substream contributes to the aggregate stream. The network masks divide the aggregate stream such that the Class A address space and Class D/E address space each correspond to a substream, and the remaining 12 streams are created by partitioning the Class B and C address along bits 2-5[3]. Although this partitioning is arbitrary, it is

[3]This partitioning was motivated by making each of the Class B and Class C streams the same "size" with respect to number of network addresses.

sufficient to give the statistical properties for substreams with a range of means.

There are several interesting observations with regard to the distribution of packets throughout the IP address space:

- Class A destinations accounts for less that 2% of the packet arrivals while consuming half of the total IP address space.
- Two of the Class C streams (i.e., those addresses in the range $192.0.0.0 - 207.255.255.255$) account for 60% of the packets but consume only $1/16$ of the IP address space.
- Half of the high order Class C (i.e., $208.0.0.0 - 223.255.255.255$) and a quarter of the high order Class B (i.e., $176.0.0.0 - 191.255.255.255$) address space had almost no arrivals. For this reason, we do not consider these substreams in the analysis that follows.

Filter Mask	2AM	9PM	3PM
$0.0.0.0 - 127.255.255.255$ (Class A)	1.6%	1.6%	1.7%
$128.0.0.0 - 135.255.255.255$ (Class B)	20%	20%	21%
$136.0.0.0 - 143.255.255.255$ (Class B)	6.9%	5.9%	3.9%
$144.0.0.0 - 151.255.255.255$ (Class B)	3.0%	3.0%	2.4%
$152.0.0.0 - 159.255.255.255$ (Class B)	4.2%	7.7%	6.3%
$160.0.0.0 - 167.255.255.255$ (Class B)	3.0%	3.0%	2.4%
$168.0.0.0 - 175.255.255.255$ (Class B)	0.6%	1.4%	1.0%
$176.0.0.0 - 183.255.255.255$ (Class B)	0.0%	0.0%	0.0%
$184.0.0.0 - 191.255.255.255$ (Class B)	0.0%	0.0%	0.0%
$192.0.0.0 - 199.255.255.255$ (Class C)	21%	26%	22%
$200.0.0.0 - 207.255.255.255$ (Class C)	40%	32%	39%
$208.0.0.0 - 215.255.255.255$ (Class C)	0.0%	0.1%	0.2%
$216.0.0.0 - 223.255.255.255$ (Class C)	0.0%	0.0%	0.0%
$224.0.0.0 - 255.255.255.255$ (Class D/E)	0.3%	0.1%	0.2%

Table 1. Network filter mask and percent of traffic for 2AM, 9PM and 3PM traces.

4.1 Substream Arrival Distribution

Figure 10 shows the empirical probability density for several of the component substreams of the 3PM trace. For clarity of presentation, we include only the substreams which compose more than 3% of the aggregate stream. The solid lines depict the empirical streams, while the dashed lines illustrate an analytical log-normal distribution whose parameters were determined using the MLE. As the figure shows, the log-normal distribution provides a good fit for the streams with a larger mean, but those with smaller means contain a large number of intervals with no arrivals, which makes the log-normal fit poor. The data suggests that light streams can not be modeled well by the log-normal distribution, at least not without correction factors such as discretizing the log-normal near zero.

4.2 Substream Correlation Structure

Figure 11 shows log-variance plots of the component substreams for each trace. For the medium and high utilization streams (i.e., 3PM and 9PM traces), the component substreams exhibit the same degree of self-similarity as the aggregate streams shown in Figure 7 (except for the class D/E traffic, which is uncorrelated for all three traces). The larger substreams of the low utilization trace (i.e., 2AM trace) also exhibits the same correlation structure as the aggregate stream. However, in the 2AM trace, the degree of self-similarity decreases with the mean of the substream. This observation is consistent with other studies [11], which determine that sample paths with low utilization exhibit a smaller degree of self-similarity than streams with high utilizations. The data in Figure 11 suggests that the degree of

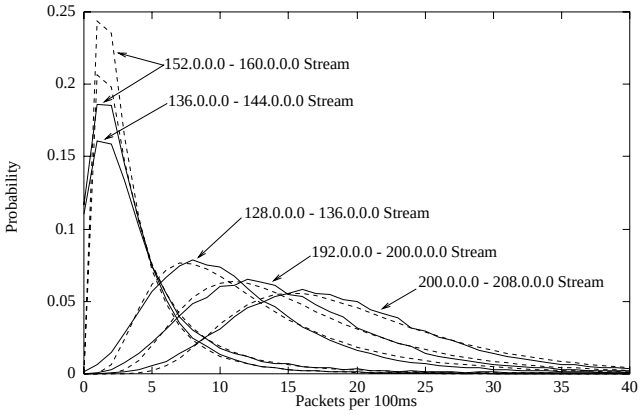


Figure 10. Packet arrival density for 3PM composite substreams.

self-similarity of substreams obtained by partitioning a synthetic aggregate stream (i.e., assigning network addresses) preserves the self-similarity as a function of the mean. This conclusion holds for the UVAnet aggregate stream, whose mean is measured in hundreds of packets per second; however, further research is needed to determine at which point (in terms of packets per second) the self-similarity is not preserved.

5 Conclusions and Future Work

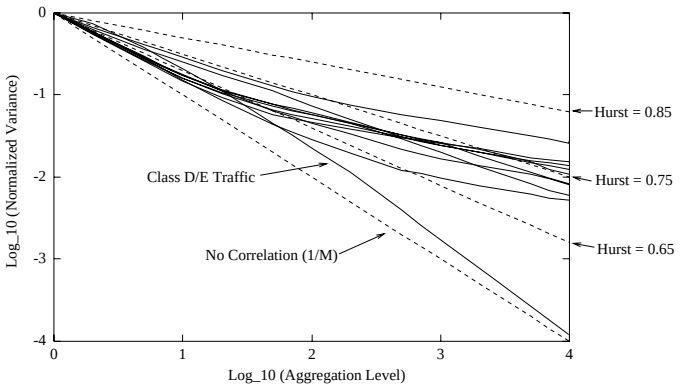
In this paper we presented the statistical characteristics of long packet traces exchanged between UVAnet, VERnet and BBNplanet. We focused on three representative 90-minute traces of packets leaving the UVA network. We first considered the distribution and correlation of packet sizes and found that the densities are nearly identical for all three traces, and are short-term correlated. Next, we considered the density and correlation structure of the arrivals for each trace. We showed that the arrival density can be modeled with a log-normal distribution and that the arrivals are self-similar, exhibiting significant long-range dependencies as found in Ethernet LAN traffic studies. Finally, we analyzed the component substreams of the aggregate traces. In our example, the component substreams which compose more than 3% of the aggregate stream are also log-normally distributed; however, the component substreams with very low packet arrivals tend to deviate from the parameters of the aggregate stream.

Future work will center on the development of an efficient analytic background traffic model for WAN simulation that will reflect the statistical properties of the empirical data. This characterization study confirms the applicability of self-similar models for modeling the aggregate packet arrival process for a large campus network such as the university campus studied. We intend to model the traffic entering a WAN backbone at a network access point with a single self-similar traffic generator, using the statistics derived in the analysis in this paper, and then to split the outgoing traffic into destination-based substreams. This approach is attractive in that a single generator per network access point (NAP) makes the computational burden of background traffic generation linear in the NAPs on the backbone.

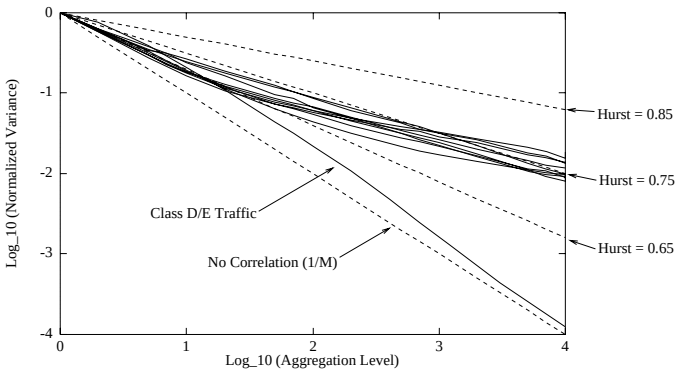
References

[1] Snoop Packet Filter. Sun Solaris 2.5 man page. Sun Microsystems, 1996.

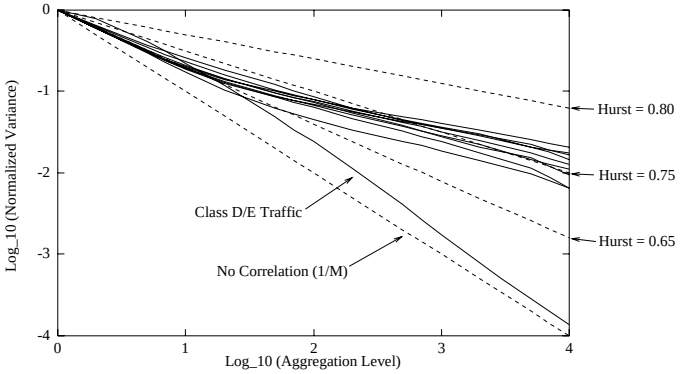
[2] UVA Packet Traces. Available via ftp at ftp.cs.virginia.edu in /pub/mtl8c.



(a) 2AM Trace - Low Utilization.



(b) 9PM Trace - Medium Utilization.



(c) 3PM Trace - High Utilization.

Figure 11. Log-variance plots of (a) 2AM, (b) 9PM and (c) 3PM component substreams.

[3] A. Adas and A. Mukherjee. On Resource Management and QoS Guarantees for Long Range Dependent Traffic. In *Proc. IEEE Infocom*, pages 779–787, April 1995.

[4] J. Beran. Statistical Methods for Data with Long-Range Dependence. In *Statistical Science, 7(4)*, pages 404–427, 1992.

[5] D. R. Cox. Long-Range Dependence: A Review. In *Statistics: An Appraisal, Proc. 50th Anniversary Conference*, pages 55–74, Ames, IA: Iowa State University Press. Iowa State Univ. Press, 1984.

[6] Mark E. Crovella and Azer Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. In *Proc. ACM Sigmetrics '96*, pages 160–169, May 1996.

[7] V. Frost and B. Melamed. Traffic Modelling for Telecommunications Networks. *IEEE Communications Magazine*, 32(3):70–81, March 1994.

[8] M.W. Garrett and W. Willinger. Analysis, Modeling and Generation of Self-Similar VBR Video Traffic. In *ACM Sigcomm 1994*, London, UK, August 1994.

[9] H. Heffes and D. M. Lucantoni. A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):856–868, September 1986.

[10] Wing-Cheong Lau, Ashok Erramilli, Jonathan L. Wang, and Walter Willinger. Self-Similar Traffic Generation: The Random Midpoint Displacement Algorithm and Its Properties. In *Proc. IEEE ICC '95*, pages 466–472, Seattle, Washington, 1995. IEEE.

[11] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, February 1994.

[12] B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian Motions, Fractional Noise and Applications. *SIAM Review*, 10:422–437, 1968.

[13] Vern Paxton. Fast Approximation of Self-Similar Network Traffic. Technical Report LBL-36750, Lawrence Berkeley Laboratory and EECS Division, University of California, Berkeley, April 1995.

[14] R.Jain and S. A. Routhier. Packet Trains: Measurements and a New Model for Computer Network Traffic. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):986–995, September 1986.