How *Real* Can Synthetic Network Traffic Be?

Félix Hernández-Campos F. Donelson Smith Kevin Jeffay

Department of Computer Science
University of North Carolina
Chapel Hill, NC 27599-3175
{fhernand, smithfd, jeffay}@cs.unc.edu
http://www.cs.unc.edu/Research/dirt

At some point, most networking researchers simulate or emulate networks in order to understand the behavior or performance of some piece of networking technology. In performing these experiments, there is a clear need to have the ability to generate realistic synthetic workloads; synthetic network traffic whose statistical properties match those of traffic observed on a real network link. In our work, we are developing a method of traffic modeling and synthetic generation that can realize a new level of realism in network experiments that is not possible with existing techniques.

Our approach follows the philosophy of using source-level descriptions of network traffic advocated by Floyd and Paxson [1]. We develop a new source-level model of the data exchange dynamics inside a TCP connection that is *application-independent*, and therefore amenable to modeling the complete mix of TCP traffic seen on a link. This is an advance over existing source-level modeling techniques that only model the traffic from a single application (*e.g.*, HTTP models [2]).

Our modeling of TCP connections distinguishes two cases. The first case is motivated by the observation that most applications communicate using a series of requests and responses. These connections can be described as a sequence of the form:

$$<(a_1, b_1, t_1), (a_2, b_2, t_2),\ldots, (a_n, b_n)>$$

where a_i and b_i represent the sizes of application-level data units flowing in the forward and the reverse directions of the connection respectively, and t_i represents the idle time between an exchange. The second case comes from connections wherein endpoints exchange data units concurrently. We represent these connections using two independent sequences of the form:

$$<(a_1, ta_1), (a_2, ta_2),\ldots, (a_n)>,\ <(b_1, tb_1), (b_2, tb_2),\ldots, (b_m)>$$

Our model, called the *a-b-t model*, is powerful enough to capture the essential nature of a connection at the source-level, and yet is simple enough to be populated directly from measurements.

We have developed a set of techniques and tools to convert an arbitrary packet header trace into a set of *connection vectors* of the forms shown above. These vectors represent the workload of TCP in terms of communication demand at the socket-level, and provide the foundation for *closed-loop* traffic generation of *entire traffic mixes*. This enables us to conduct experiments in which some network mechanism (*e.g.*, a TCP variant, an AQM scheme, *etc.*) is exposed the full range of source-level behaviors observed in one (or more) real network links.

The connection vectors are input to synthetic traffic generators we have developed for use in network simulators and testbeds. The most straight-forward traffic generation technique is to directly *replay* the source-level behavior captured by a set of connection vectors. In this case, each measured connection vector results in one synthetically generated connection with the relative start times between connections preserved. We are also investigating techniques for re-sampling and sub-sampling of connection vectors in order to derive new source-level traces that preserve the statistical variability of the original trace but result in different (and controllable) levels of load when used in a replay.

The direct replay of connection vectors has an important benefit: it enables us to compare the properties of a real trace and the ones of its source-level replay. This comparison provides not only a way of validating our modeling technique, but also a method for studying to the extent to which synthetic traffic is *realistic*, *i.e.*, the extent to which it can approximate the measured properties of real traffic. Our goal is to demonstrate the potential of our approach, and expose the limitations of synthetic traffic generation.

We are currently involved in a comprehensive experimental study to develop methods for tuning network-dependent parameters of the experimental environment (*e.g.*, round-trip times, window sizes, access bandwidths) to control the degree of realism in the synthetically generated traffic. Ultimately, the goal is to approach a high-fidelity reproduction of traffic observed on a network link. Our metrics for evaluating the degrees of realism in synthetic traffic include the following:

- The link load or throughput – the number of bits per second (including protocol headers) transmitted on a link.
- The statistical properties of the time series of counts of arriving packets and bytes on a link in an interval of time (*e.g.*, Hurst parameter estimates, wavelet spectra).
- The number of active TCP connections over an interval of time (typically one second).
- The distributions of TCP connections durations and rates.
- The distribution of packet sizes on the link.

Our cases studies include traces from several Internet links, including an OC-48 Abilene backbone link and a 1 Gbps Ethernet link connecting the UNC campus with its Internet service provider. The early results are promising. We have been able to demonstrate that through judicious control of network-dependent parameters, high-fidelity reproduction of traces of traffic from these links is possible in a network testbed.

REFERENCES

[1] S. Floyd, and V. Paxson, Difficulties in Simulating the Internet, *IEEE/ACM ToN*, vol. 9, num. 4, August 2001, pp. 392–403.

[2] P. Barford, and M. Crovella, Generating Representative Web Workloads for Network and Server Performance Evaluation, Proc. ACM SIGMETRICS, July 1998.