# Networking Lab

## Understanding Behavior and Control of the Internet

**by Steven H. Low**

**The Internet has become, in a short span of four decades, critical infrastructure for the world and a common platform for innovation with impact far beyond communications.** It connects 20% of humanity and offers a trillion websites as of 2008. Underlying the apparent convenience and simplicity, however, is a complex array of hardware, software, and control algorithms. Indeed the Internet is the largest distributed nonlinear feedback control system ever built. Despite the Internet's importance, there is no fundamental understanding of its global behavior and future evolution. Will congestion-induced collapse, first detected on the Internet in 1986 with only 5,000 hosts, reoccur in the modern network with more than half a billion hosts? How can we optimize the interaction between congestion control, routing, and the wireless infrastructure? Will performance and stability problems arise as the Internet continues to scale in size, speed, scope, heterogeneity and complexity?

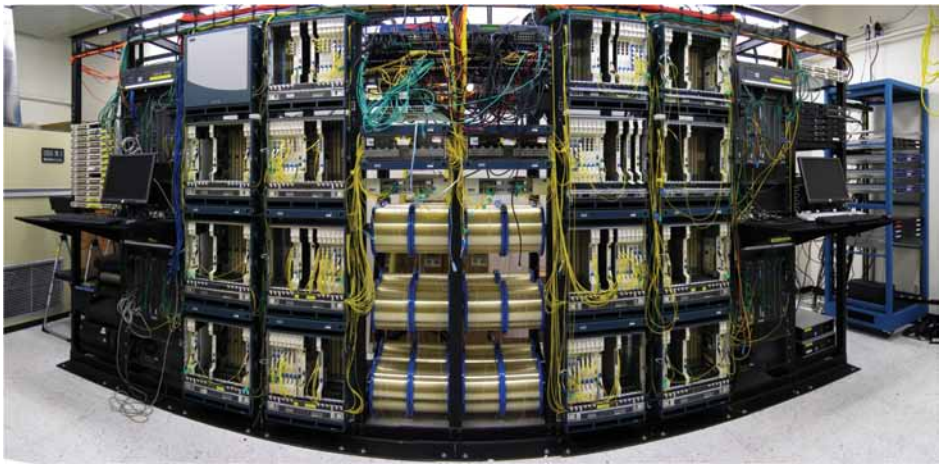

Figure 1: Caltech WAN-in-Lab testbed (2008).

Although the working elements (hardware, software, and algorithms) that make up the Internet are well understood, their collective behavior is much harder to quantify, and is one of the defining features of a complex system. A comment made by Caltech's Gordon and Betty Moore Professor, Emeritus, Carver Mead, some 20 years ago about computational neural systems applies equally well to the Internet today: "The complexity of a computational system derives not from the complexity of its component parts, but rather from the multitude of ways in which a large collection of these components can interact. Even if we understand in elaborate details the operation of every nerve channel and every synapse, we will not by so doing have understood the neural computation."

Developing a fundamental understanding of large-scale networks remains a most critical and difficult challenge in networking research. It is this interconnection that provides the myriad of communication services that we now take for granted, but that also allows the failure of a single power plant to black out a continent or the crash of a stock to trigger a regional financial meltdown. A comprehensive theory of large-scale networks is the only certain way to harvest their power and manage their risk.

Since 2000, Caltech's Networking Lab has established an integrated research program where theory, experiment, and networking infrastructure can intimately inform and influence each other. We have been developing a mathematical theory of the Internet, focusing particularly on congestion control. One outcome has been the invention of a new congestion control algorithm called FAST TCP. Moreover, we have built a unique testbed (WAN-in-Lab, see Figure 1) to test our theory, and have also started a company FastSoft to drive the deployment of FAST TCP. Our research program has had two main results. First, it has firmly established a theoretical basis for TCP (Transmission Control Protocol) design, and has brought mathematical rigor to TCP research. Second, it has demonstrated the feasibility of scaling TCP to very high speed global networks of the future and has developed some mathematical tools and design approaches to develop and evaluate new protocols.

## Theory

With Professor John Doyle, we have developed a theory of Internet congestion control based on convex optimization and control theory[1, 2, 5]. In this theory, congestion control is a distributed asynchronous primal-dual algorithm carried out over the Internet in real-time to solve an abstract global optimization problem. Different TCP (Transmission Control Protocol) algorithms differ merely in the objective functions they implicitly optimize. This theory allows us to understand the limitations of the current TCP and to design new algorithms. Until recently, the majority of TCP research has largely been simulation-based, often using only a single bottleneck link and a single class of algorithms. Our theory can predict the equilibrium behavior of an arbitrary network operating under any TCP-like algorithms. Moreover, for the first time, we can predict and design efficiency and stability properties in the presence of feedback delay of arbitrary networks.

The theory has also led to the discovery of some counterintuitive behaviors of large-scale networks. For instance, a network with heterogeneous congestion control algorithms can have multiple equilibrium points, and not all of them can be locally stable. Such phenomena will become more important as the Internet becomes more heterogeneous. It is generally believed that a resource allocation policy can either be fair or efficient but not both. We characterize exactly the tradeoff between fairness and throughput in general networks. The characterization allows us both to produce the first counter-example and trivially explain all the previous supporting examples in the literature. Surprisingly, our counter-example has the property that a fairer allocation is always more efficient. Intuitively, we might expect that increasing link capacities always raises aggregate throughput. We show that not only can throughput be reduced when some link increases its capacity, but it can also be reduced when all links increase their capacities by the same amount. Often, interesting and counterintuitive behaviors arise only in a network setting where flows interact through shared links in intricate and surprising ways. Such behaviors are absent in single-link models that were prevalent in the TCP literature; and they are usually hard to discover or explain without a fundamental understanding of the underlying structure. Given the scale and diversity of the Internet, it is conceivable that such behaviors are more common than we realize, but remain difficult to measure due to the complexity of the infrastructure and our inability to monitor it closely. A mathematical framework thus becomes indispensable in exploring structures, clarifying ideas, and suggesting directions. We have demonstrated experimentally some of these phenomena using the WAN-in-Lab testbed.

## Experiment

We have implemented the insights from theoretical work in a new protocol FAST TCP[4] and have worked with our collaborators to test it in various production networks around the world[3]. Physicists, led by Professor Harvey Newman of Caltech's Division of Physics, Mathematics and Astrnomy, have been using FAST TCP to break world records on data transfer from 2002–2006. Perhaps more important than the speeds and data volumes of these records, has been the impact of our experiments on the research and application of TCP. For instance, the Internet2's Land Speed Record in late 2002 was attained in an experiment where data was transferred for only 16 seconds—the techniques that were used in that, and all previous, records, did not address protocol issues and were therefore so fragile that high throughput could be sustained only momentarily. As soon as network fluctuations occurred, throughput would collapse, and the experiment was terminated. Our experiments in the 2002 SuperComputing Conference were the first to sustain line rate (at mutlti-Gbps) for several hours, using a standard packet format. At that time, the high performance computing community was very skeptical about the feasibility of using TCP as the transport protocol for high-speed (multi-Gbps) long-distance (4,000 miles) transfer of large volumes (petabytes, or $10^{15}$ bytes) of scientific data. Our experiments settled the debate and changed the direction in the high-energy physics community from developing

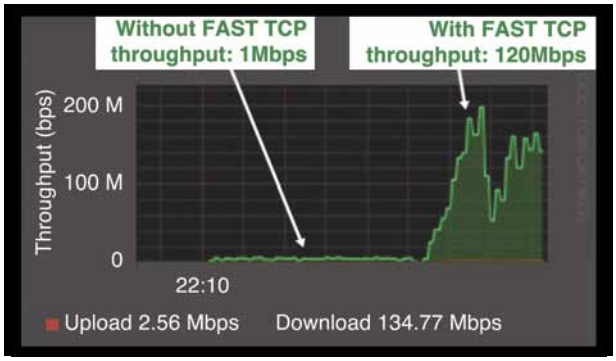


Figure 2: In June 2007, part of Sprint network suddenly suffered up to 15% packet loss for about 15 minutes. During this period, the throughput between FastSoft's San Francisco data center and New York data center dropped to 1 Mbps. FastSoft acceleration provided extreme resilience to packet loss and restored the throughput to 120 Mbps.

UDP-based data transfer solutions to TCP-based solutions. This and subsequent experiments have been widely reported in the media including *Nature, National Geographic, New Scientist, The Economist, BBC, CNN, Reuters, Business Week, The Australian, MSNBC, Yahoo! News, PC Magazine, ComputerWorld, CNet,* etc.

## Testbed

We have built a one-of-a-kind academic testbed, WAN-in-Lab, at Caltech for the design, development, and evaluation of high speed network protocols (see Figure 1). It uses real carrier-class networking hardware to avoid the artifacts introduced by network simulation and emulation, while being localized to allow detailed measurement of network behavior. WAN-in-Lab, funded by the Lee Center, NSF, ARO, Cisco and Corning, is literally a wide-area-network—with 2,400 km of long-haul fibers, MEMs optical switches, wavelength division multiplexing equipment such as optical amplifiers and dispersion compensation modules, routers, servers and accelerators. It provides a unique platform where new algorithms can be developed, debugged, and tested first on WAN-in-Lab before they are deployed in the field, thus shortening the cycle of design, development, testing and deployment.

## Deployment

We have spun off a company, FastSoft, in 2006 to drive the development and deployment of FAST TCP. FAST TCP forms the core of a set of web acceleration technologies developed at FastSoft for the delivery of web applications and contents over the Internet to enterprises and consumers. Such technologies will become increasingly important with the proliferation of video, dynamic content, and cloud services. They are accelerating the world's second largest content distribution network and other Fortune 100 companies, typically by 1.5 to 30 times. LEE

*Steven H. Low is Professor of Computer Science and Electrical Engineering.*

**Read more at:** http://netlab.caltech.edu

### References

[1] S. H. Low, "A Duality Model of TCP and Queue Management Algorithms," *IEEE/ACM Trans. On Networking*, **11**(4):525-536, August 2003.

[2] F. Paganini, Z. Wang, J. C. Doyle, and S. H. Low, "Congestion Control for High Performance, Stability and Fairness in General Networks," *IEEE/ACM Trans. on Networking,* **13**(1):43–56, February 2005.

[3] C. Jin, D. X. Wei, S. H. Low, G. Buhrmaster, J. Bunn, D. H. Choe, R. L. A. Cottrell, J. C. Doyle, W. Feng, O. Martin, H. Newman, F. Paganini, S. Ravot, and S. Singh, "FAST TCP: From Theory to Experiments," *IEEE Network*, **19**(1):4-11, January/February 2005.

[4] D. X. Wei, C. Jin, S. H. Low, and S. Hegde, "FAST TCP: Motivation, Architecture, Algorithms, Performance," *IEEE/ACM Transactions on Networking*, **14**(6):1246-1259, December 2006.

[5] Mung Chiang, Steven H. Low, A. Robert Calderbank, and John C. Doyle, "Layering as Optimization Decomposition: A Mathematical Theory of Network Architectures," *Proceedings of the IEEE*, **95**, 255-312, January 2007.