This project investigates issues in "large-scale" networks. Here "large-scale" refers to networks with large number of high capacity nodes and transmission links, and shared by a large number of users. It is almost impossible to simulate such networks and we aim to develop mathematical tools to uncover their structure, illuminate issues, and design solutions.

In the three year project period, we have developed a mathematical theory to understand large scale networks under end-to-end control such as the Internet. We have used the theory to explain the deficiencies of the current TCP congestion control algorithm and design new ones. We have implemented the new design as a new protocol FAST TCP. We have worked with a large number of collaborators around the world to test, and refine, the protocol on production networks. The software prototype that we developed has been used by the high energy physics community to break world records on data transfer in the 2003, 2004, and 2005. We have just started a company to transfer the technology to the industry.
AFOSR Final Performance Report

Large scale self-organizing information distribution system

PI: Steven Low
Institute: California Institute of Technology
1200 E. California Blvd, Pasadena, CA 91125
Award number: F49620-03-1-0119
Award period: Jan 1, 2003 – Dec 31, 2005
Award: $300,000

Final report period: Jan 2003 – Dec 2005

Summary
My Networking Lab at Caltech conducts research in control and optimization of networks and protocols. I have built a research program that integrates theory, algorithm, implementation, experiments, and applications, where these elements inform and influence each other intimately. This project is part of our overall program with a broad objective of investigating issues in “large-scale” networks. Here “large-scale” refers to networks with large number of high capacity nodes and transmission links, and shared by a large number of users. It is almost impossible to simulate such networks and we aim to develop mathematical tools to uncover their structure, illuminate issues, and design solutions.

The focus of our study has slightly changed since the submission of the proposal. Even though the new focus still studies issues in distributing large contents over “large-scale” networks, we have decided to postpone the study of algorithms for server allocation and placement. We have found that before we can optimally deploy servers around the network, we first need to develop means for them to transfer large amount of data at high speed over large distance. This is the objective of the Caltech FAST project, of which this is a part.

In the three year project period:
1. We have developed a mathematical theory to understand large scale networks under end-to-end control such as the Internet.
2. We have used the theory to explain the deficiencies of the current TCP congestion control algorithm and design new ones.
3. We have implemented the new design as a new protocol FAST TCP.
4. We have worked with a large number of collaborators around the world to test, and refine, the protocol on production networks. The software prototype that we developed has been used by the high energy physics community to break world records on data transfer in the 2003, 2004, and 2005.
5. We have just started a company to transfer the technology to the industry.

In the following, we describe these activities and the major findings in more detail.
Motivation

The high energy and nuclear physics (HENP) community is exploring the fundamental interactions, structures and symmetries that govern the nature of matter and spacetime in the Universe. The largest HENP experiments are those in the late stages of preparing to take data at CERN's Large Hadron Collider beginning in 2007; these collaborations number around 3000 physicists and engineers from 200 universities and laboratories around the world. One of the principal tenets of the HENP community is the principal that all physicists should have the tools and access to the experiment's data, no matter where they are located. Thus, the community is dependent on the availability of high performance networks which are used to move scientific datasets both from the experiment itself to remote locations, and, in the case of result and Monte Carlo simulation data, between pairs of institutes worldwide. Rapid and reliable data transport, at speeds from 1Gbps through 10Gbps, then to 100Gbps in the next few years, is a key capability for the community.

As the scientists build up their network infrastructure, they immediately hit the bottleneck imposed by the TCP algorithm. As their network grows in size and capacity, the efficiency of the network drops steadily. One of the causes has been identified early to be TCP algorithm, and many ad hoc tweaks have been proposed and tried to address the problem, without much success. This motivates a more comprehensive approach of the FAST Project.

Theory, algorithm, implementation, experiments

We have developed a duality model that interprets any TCP/AQM algorithm as a distributed asynchronous primal-dual algorithm carried out over the Internet in real-time in the form of congestion control to solve a utility maximization problem. Different algorithms differ merely in the utility functions they implicitly optimize. The model allows us to understand the limitations of the current TCP, and the many hacks that have been proposed, and design new algorithms. Until five or so years ago, the state of the art in TCP research has been simulation-based using simplistic scenarios, with often a single bottleneck link and a single class of algorithms. We have now a theory that can predict the equilibrium behavior of an arbitrary network under any TCP-like algorithm. Moreover, for the first time, we can prove, and design, their stability properties in the presence of feedback delay for arbitrary networks. We have developed new theoretical tools for proving global and local stability in the presence of feedback delay.

We have implemented the insights from this series of theoretical work in a software prototype FAST TCP. We have been working with our collaborators worldwide in the last three years to test and refine the algorithm using global production networks, including Abilene (Internet2 backbone), CENIC, HENP trans-Atlantic network, etc. The HENP community has been using our software to break world records in data transfer in 2003, 2004, and 2005. For example, in the last SuperComputing Conference in November 2005, Caltech HENP professor Harvey Newman and his team used FAST
TCP in their experiment that set a record of 150Gbps aggregate transfer rate over a collection of 10Gbps links.

Figure 1 below shows the network diagram for the Bandwidth Challenge at SuperComputing Conference in November 2004.

Figure 1. Network diagram for 2004 SuperComputing Conference Bandwidth Challenge (Source: Professor Harvey Newman, Caltech). The record-breaking experiment utilizes Abilene (Internet2 backbone and the trans-Atlantic high-energy physics network between Chicago PoP and CERN in Geneva.

Figure 2 below shows some experimental results using FastTCP. The left panel shows data transfer across the Atlantic from Sunnyvale through Chicago to Geneva: FastTCP improved the throughput by 5 times. The right panel compares the performance of FastTCP with several other TCP variants over an emulated lossy network. It shows the throughput achieved by these TCP variants as a function of random packet loss rate: FastTCP was able to achieve close to the theoretical maximum while other variants collapse when the loss rate exceeds 5%.
Technology transfer
We have started a company, FastSoft, Inc., to commercialize the research prototype. The first product of the company will be an appliance, FastSoft Aria 100, that embodies the basic FastTCP technology. It improves TCP transfer dramatically – by 30 times in "typical" conditions – and robustly – across a wide range of operating conditions. Figure 3 shows a testbed we use to benchmark the appliance. It uses an emulated network to emulate various operating conditions over WAN. We measured the throughput achieved at the application layer (by iperf) between the sender and receiver, with and without Aria 1000 in front of the sender.
Figure 3. Testbed for benchmark of Aria 1000. The emulated network emulate a WAN with different capacity, distance of transfer, and random packet loss rate. The sender and the receiver are Linux servers.

Figures 4(a) and 4(b) compare the throughput in various operating conditions without Aria 1000 (left panel) and with Aria 100 (right panel). As shown in the figures, Aria 1000 not only achieves a much higher throughput, more important, the performance is much more predictable.

Figure 4(a). Throughput comparison over a WAN with 3Mbps capacity. Left: without Aria 1000; right: with Aria 1000.

Figure 4(b). Throughput comparison over a WAN with 155Mbps (OC3) capacity. Left: without Aria 1000; right: with Aria 1000.

**Personnel supported**
The following people have been supported by this award in the last 3 years:

- Other Faculty Salaries category: Joon-Young Choi (12/1/04 to 2/28/05)
- Bi-Weekly Staff category:
In addition, the grant has also supported the following students in some of their travels to technical conferences: Jiantao Wang, Hyojeong Choe, Martin Suchara.

Publications
In the last 3 years, 20 journal papers/book chapters and 20 conference papers have been published or accepted for publication on the FAST project. The journal papers/book chapters are listed below. Almost all of the conference papers evolve into a journal paper and therefore they are not separately listed.

Journal papers and book chapters

Interactions/transitions
We have developed close collaborative relations with applications community as well as the mathematics experts. On the theory side, we are working closely with

- Prof John Doyle of EE/CDS/BE Departments, Caltech
- Prof Fernando Paganini of EE Department, UCLA.
On experiment and applications side, we are working closely with

- Prof Harvey Newman of Physics Division, Caltech
- Jim Pool, CACR, Caltech
- Les Cottrell of SLAC, Stanford University
- Guy Almes of Internet2 (now NSF)
- David Lapsley of Haystack Observatory, MIT
- Olivier Martin of CERN
- Linda Winkler, TeraGrid, Argonne National Lab
- Bob Aiken, Chris McGugan, Steven Yip, Cisco
- Jim Grey, Microsoft
<table>
<thead>
<tr>
<th>PI Name</th>
<th>P.I. Organization</th>
<th>AFSOR PM Name</th>
<th>Title of Research Effort</th>
<th>Follow-on Use Category</th>
<th>Contact Person Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steven Low</td>
<td>Caltech</td>
<td>Dr. Robert L. Herkowitz</td>
<td>Large scale self-organizing information distribution system</td>
<td>Technology Transfer</td>
<td>FastSoft, Inc.</td>
</tr>
<tr>
<td>Description of Results</td>
<td>Application or Potential Application</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FastSoft, Inc is a startup company that has licensed the technology developed in this and related projects from Caltech as the basis of a commercial product.</td>
<td>To accelerate data transfer over the Internet.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FastSoft is currently privately funded. Significant investment from the private sector has gone into taking the basic research and prototype and develop it into industrial grade enterprise product. This is only the first year of the company, so no official product release yet.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>